

THE INTERNATIONAL HANDBOOK OF SPACE TECHNOLOGY

Malcolm
Macdonald

Viorel
Badescu

Editors



Springer Praxis Books

Astronautical Engineering

For further volumes:
<http://www.springer.com/series/5495>

Malcolm Macdonald
Viorel Badescu
Editors

The International Handbook of Space Technology

 Springer

Published in association with
Praxis Publishing
Chichester, UK

 PRAXIS

Editors

Malcolm Macdonald
University of Strathclyde
Glasgow
Scotland

Viorel Badescu
Candida Oancea Institute
Polytechnic University of Bucharest
Bucharest
Romania

ISBN 978-3-642-41100-7 ISBN 978-3-642-41101-4 (eBook)

DOI 10.1007/978-3-642-41101-4

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013958135

© Springer-Verlag Berlin Heidelberg 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*For those who know that great thoughts reduced to practice
become great acts*

Foreword

I believe strongly in the value of expanding the scope and scale of human consciousness via the permanent extension of humanity to multiple planets. This is also vitally important in preserving the long-term security of life as we know it against a natural or man-made extinction event. These reasons are why I created SpaceX and why I believe the space industry is so important.

In 2010, SpaceX became the first private company to bring a spacecraft back from orbit when our Dragon vehicle completed two revolutions of the Earth before returning safely with a splashdown approximately 800 km west of California. I am pleased to say that Dragon was in such good shape that it could be flown again just by repacking the parachutes.

That was the first of many steps toward what I believe to be the utterly critical goal of orbital space flight: complete and rapid reusability. This is the fundamental breakthrough in space flight that must be achieved, whether by SpaceX or another company. Our Falcon 9 launch vehicle, for example, costs over \$50M to build, but the propellant (liquid oxygen and kerosene) cost is only \$0.2M, implying that more than a 100-fold improvement in the cost of space flight is achievable. This ratio is comparable to that of aircraft, which are, of course, highly reusable. The quest for this breakthrough is what makes this handbook on space technology so valuable at this time, contributing, as it will, to the proliferation and exploitation of space technology to enable the consciousness of humanity to move beyond our cradle and to explore beyond this Earth.

The initiative taken by Springer Praxis to initiate this handbook is warmly welcomed, while the efforts of the editorial team of Malcolm Macdonald and Viorel Badescu in realizing their vision over 25 chapters, written by experts from all around the world, are commended.

Wherever in the world they may find themselves, I hope the readers of this handbook will use the knowledge contained within to devote their energy toward achieving great breakthroughs.

Elon Musk

Acknowledgments

A critical part of writing any book is the review process, and the authors and editors are very much obliged to the following researchers who patiently helped them read-through the subsequent chapters and who made valuable suggestions:

Erik Andrews	Back Nine Engineering, Laurel, Maryland, USA
Viorel Badescu	Polytechnic University of Bucharest, Bucharest, Romania
Tim Barfoot	University of Toronto, Toronto, Canada
Max Calabro	The Inner Arch, Villennes, France
Daniel Choukroun	Delft University of Technology, Delft, The Netherlands
Alan P. Cudmore	Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA
George Dakermanji	Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA
Laurent Dala	University of the Witwatersrand, Johannesburg, South Africa
Chris Damaren	University of Toronto, Toronto, Ontario, Canada
Daniel F. DiFonzo	RaySat Antenna Systems/Spacenet Integrated Government Solutions, McLean, Virginia, USA
Jerry Fiedziuszko	Space Systems/Loral, Fabyan Way, Palo Alto, California, USA
James Haines	European Space Research and Technology Centre, European Space Agency (ESA-ESTEC), Noordwijk, The Netherlands
Hirohisa Hara	National Astronomical Observatory of Japan (NAOJ), Mitaka, Tokyo, Japan
Michael A. Johnson	Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA
Robert Kok	McGill University, Montréal, Québec, Canada
George J. Komar	Earth Science Technology Office, Science Mission Directorate, National Aeronautics and Space Administration (NASA) Headquarters, Washington, District of Columbia, USA
Chandra Kudsia	University of Waterloo, Waterloo, Ontario, Canada
James R. Lemen	Lockheed Martin Solar and Astrophysics Laboratory, Palo Alto, California, USA
Roger Longstaff	Reaction Engines Ltd., Abingdon, England
James P. Lux	Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, California, USA
Giorgio Magistrati	European Space Research and Technology Centre, European Space Agency (ESA-ESTEC), Noordwijk, The Netherlands
Malcolm Macdonald	Advanced Space Concepts Laboratory, Strathclyde Space Institute, University of Strathclyde, Glasgow, Scotland
Landis Markley	Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA

Colin McInnes	Advanced Space Concepts Laboratory, Strathclyde Space Institute, University of Strathclyde, Glasgow, Scotland
Robert Moore	The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA
Sa'id Mosteshar	London Institute of Space Policy and Law, London, England
Zeina Mounzer	Telespazio VEGA Deutschland GmbH, Darmstadt, Germany
Mark Nelson	Institute of Ecotechnics, London, England
Olivier Notebaert	Astrium Satellites SAS, Toulouse, France
David Patterson	European Space Operations Centre, European Space Agency (ESA-ESOC), Darmstadt, Germany
Antoine Provost-Grellier	Thales Alenia Space, Cannes, France
Abbas Salim	Lockheed Martin (Ret.), Sunnyvale, California, USA
Bruce Savadkin	Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA
Julian Santiago Prowald	European Space Research and Technology Centre, European Space Agency (ESA-ESTEC), Noordwijk, The Netherlands
Stephen A. Shinn	Flight Projects Directorate, Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, Maryland, USA
David J. Southwood	Imperial College, London, England
Arthur Smith	Fluid Gravity Engineering, Emsworth, England
Alejandro Torres	IberEspacio, Madrid, Spain
Richard Tremayne-Smith	Consultant, Ash Vale, England
Dominique Valentian	SAFRAN/Snecma, Paris, France
Angelos Vourlidis	United States Naval Research Laboratory, Washington, District of Columbia, USA
Joanne Wheeler	CMS Cameron McKenna LLP, London, England
Jim Wood	Department of Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow, Scotland
Amir I. Zaghoul	Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

The editors, furthermore, owe a debt of gratitude to all authors. The development of this handbook has been a long process and has endured earthquakes, building fires, tsunamis, volcanic eruptions, nuclear disaster, and the largest asteroid impact in a century; however, collaborating with these stimulating colleagues has always been a privilege and a satisfying experience. We also wish to thank strongly Ronald Rapp, who initiated the project, David M. Harland for his thorough editorial assistance, Ruaridh Clark for his assistance in implementing David's comments, Jim Wilkie for his cover design, all those at Praxis/Springer who have helped realize this book, and to wish Clive Horwood a long and happy retirement.

Contents

1	Introduction	1
	Malcolm Macdonald, Pat Norris and David B. Spencer	
2	A System-Level View of Space Projects	25
	Malcolm Macdonald	
3	Space Environments and Survivability	37
	Henry B. Garrett	
4	Introduction to Astrodynamics	61
	Malcolm Macdonald	
5	Introduction to Atmospheric Transit	99
	Richard Brown, Tom Scanlon and Jason Reese	
6	Payload Design and Sizing	117
	David Alexander and Neil Murphy	
7	Space Systems Engineering	143
	Vincent L. Pisacane	
8	Launch Systems	165
	Christophe Bonnal, Alessandro Ciucci, Michael H. Obersteiner and Oskar Haidn	
9	Structure, Mechanisms and Deployables	197
	Gerard Miglioreno and Torben K. Henriksen	
10	Electrical Power	249
	Mukund R. Patel	
11	Spacecraft Propulsion	279
	Claudio Bruno	
12	Attitude and Orbit Control Systems	323
	Bong Wie, Vaios Lappas and Jesús Gil-Fernández	
13	Thermal Systems	371
	José Meseguer, Isabel Pérez-Grande, Angel Sanz-Andrés and Gustavo Alonso	

14	Communications Systems	397
	Ali Atia and Huiwen Yao	
15	On-Board Data Systems	441
	Torbjörn Hult and Steve Parkes	
16	Flight Software	471
	Christopher Krupiarz, Annette Mirantes, Doug Reid, Adrian Hill and Roger Ward	
17	Habitation in Space	493
	Masamichi Yamashita and Raymond M. Wheeler	
18	Entry, Descent and Landing Systems	515
	Steve Lingard and John Underwood	
19	Space Robotics	541
	Kazuya Yoshida, Dragomir Nenchev, Genya Ishigami and Yuichi Tsumaki	
20	Ground Segment	575
	Richard Lowe, Dan Kent, Paul Coutinho and Kevin Halsall	
21	Technology Management	599
	Gregory L. Davis, Raphael R. Some and Andrew A. Shapiro	
22	Project Management: Relationship Between the Project Manager and the Technologist	619
	Robert J. Menrad and George W. Morrow	
23	Legal and Regulatory Issues	657
	Tanja Masson-Zwaan and Richard Crowther	
24	Advanced Concepts	677
	Les Johnson and Jack Mulqueen	
25	Mission and System Design	685
	Massimiliano Vasile, Stephen Kemble, Andrea Santovincenzo and Mark Taylor	
	Index	715

Contributors

David Alexander Rice Space Institute, Rice University, Houston, TX, USA

Gustavo Alonso E.T.S.I. Aeronáuticos, Universidad Politécnica de Madrid, Madrid, Spain

Ali Atia Orbital Sciences Corporation, Dulles, VA, USA

Christophe Bonnal Centre National d'Études Spatiales, Paris, France

Richard Brown Department of Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow, Scotland

Claudio Bruno Thermal and Fluid Sciences, United Technologies Research Center, East Hartford, CT, USA

Alessandro Ciucci European Space Agency, Paris, France

Paul Coutinho Telespazio VEGA UK Ltd., Luton, UK

Richard Crowther UK Space Agency, Swindon, UK

Gregory L. Davis Jet Propulsion Laboratory, Mechanical Systems Division, California Institute of Technology, Pasadena, CA, USA

Henry B. Garrett Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Jesús Gil-Fernández GNC Division, GMV, Tres Cantos, Madrid, Spain

Oskar Haidn Institute of Flight Propulsion, Technische Universität München, Garching, Germany

Kevin Halsall Telespazio VEGA UK Ltd., Luton, UK

Torben K. Henriksen Department of Mechanical Engineering, European Space Research and Technology Centre, European Space Agency, Noordwijk, The Netherlands

Adrian Hill Applied Physics Laboratory, Embedded Applications Group, Space Department, Johns Hopkins University, Laurel, MD, USA

Torbjörn Hult RUAG Space, Göteborg, Sweden

Genya Ishigami (石上玄也) Department of Mechanical Engineering, Keio University, Yokohama, Japan

Les Johnson Marshall Space Flight Center, National Aeronautics and Space Administration, Huntsville, AL, USA

Stephen Kemble Airbus Defense & Space, Stevenage, UK

Dan Kent Telespazio VEGA UK Ltd., Luton, UK

Christopher Krupiarz Applied Physics Laboratory, Embedded Applications Group, Space Department, Johns Hopkins University, Laurel, MD, USA

Vaios Lappas Department of Electronic Engineering, Surrey Space Centre, University of Surrey, Surrey, UK

Steve Lingard Vorticity Ltd., Chalgrove, UK

Richard Lowe Telespazio VEGA UK Ltd., Luton, UK

Malcolm Macdonald Advanced Space Concepts Laboratory, Strathclyde Space Institute, University of Strathclyde, Glasgow, Scotland

Tanja Masson-Zwaan International Institute of Air and Space Law, Leiden University, Leiden, The Netherlands

Robert J. Menrad Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, USA

José Meseguer E.T.S.I. Aeronáuticos, Universidad Politécnica de Madrid, Madrid, Spain

Gerard Migliorero Department of Mechanical Engineering, European Space Research and Technology Centre, European Space Agency, Noordwijk, The Netherlands

Annette Mirantes Applied Physics Laboratory, Embedded Applications Group, Space Department, Johns Hopkins University, Laurel, MD, USA

George W. Morrow Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, USA

Jack Mulqueen Marshall Space Flight Center, National Aeronautics and Space Administration, Huntsville, AL, USA

Elon Musk SpaceX, Hawthorne, CA, USA

Neil Murphy Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Dragomir N. Nenchev (金宮好和) Department of Mechanical Systems Engineering, Tokyo City University, Tokyo, Japan

Pat Norris Space, Defence and National Security, CGI, Leatherhead, UK

Michael H. Obersteiner Airbus Defence & Space, Bremen, Germany

Steve Parkes School of Computing, University of Dundee, Dundee, UK

Mukund R. Patel U.S. Merchant Marine Academy, Kings Point, NY, USA

Isabel Pérez-Grande E.T.S.I. Aeronáuticos, Universidad Politécnica de Madrid, Madrid, Spain

Vincent L. Pisacane United States Naval Academy (ret.), Annapolis, MD, USA

Jason Reese School of Engineering, University of Edinburgh, Edinburgh, Scotland

Doug Reid Applied Physics Laboratory, Space Department, Johns Hopkins University, Laurel, MD, USA

Andrea Santovincenzo European Space Research and Technology Centre, European Space Agency, Noordwijk, The Netherlands

Angel Sanz-Andrés E.T.S.I. Aeronáuticos, Universidad Politécnica de Madrid, Madrid, Spain

Tom Scanlon James Weir Fluids Laboratory, Department of Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow, Scotland

Andrew A. Shapiro Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Raphael R. Some Jet Propulsion Laboratory, Autonomous Systems Division, California Institute of Technology, Pasadena, CA, USA

David B. Spencer Department of Aerospace Engineering, The Pennsylvania State University, State College, PA, USA

Mark Taylor Surrey Satellite Technology Ltd., Surrey, UK

Yuichi Tsumaki (妻木勇一) Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan

John Underwood Vorticity Ltd., Chalgrove, UK

Massimiliano Vasile Advanced Space Concepts Laboratory, Strathclyde Space Institute, University of Strathclyde, Glasgow, Scotland

Roger Ward SciSys Ltd., Bristol, UK

Raymond M. Wheeler Surface Systems Division, John F. Kennedy Space Center, National Aeronautics and Space Administration, Orlando, FL, USA

Bong Wie Department of Aerospace Engineering, Asteroid Deflection Research Center, Iowa State University, Ames, IA, USA

Huiwen Yao Orbital Sciences Corporation, Dulles, VA, USA

Masamichi Yamashita (山下 雅道) Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency, Sagami, Japan

Kazuya Yoshida (吉田和哉) The Space Robotics Laboratory, Department of Aerospace Engineering, Tohoku University, Sendai, Japan

Malcolm Macdonald, Pat Norris and David B. Spencer

If I had asked people what they wanted, they would have said faster horses.

—Henry Ford (1863–1947).

As the births of living creatures at first are ill-shapen, so are all innovations, which are the births of time.

—Sir Francis Bacon (1561–1626).

Humanity has dreamt of traveling beyond our cradle and into space since, at least, the time of the Roman conquest of Greece, after the Battle of Corinth in 146 BC. In what is considered the earliest known fiction about travel to outer space, alien life-forms and interplanetary warfare, Lucian of Samosata wrote a satirical piece in 150 AD called *True History*. In part of this story a company of adventuring heroes are swept upwards in a giant waterspout shortly after sailing westward through the Pillars of Hercules (today known as the Strait of Gibraltar). After seven days and seven nights, they arrive on the Moon to find themselves embroiled in a war between the King of the Moon and the King of the Sun over the Morning Star (Venus). The Sun wins.

From then until Jules Verne's novel *From Earth to the Moon* (French: *De la Terre à la Lune*, 1865; published 1867 in English) many fanciful tales were told of humanity's travels beyond Earth. Perhaps one of the more curious of these tales was given by Cyrano de Bergerac, in 1657, in which he reasoned that the dew on morning grass

disappearing in daytime meant that the Sun must 'suck' up the dew. Hence, should enough dew be collected and stored in bottles attached to the traveler he need only stand on his lawn in a fine morning and the Sun will 'suck' him up along with the morning dew.

To this day space technology maintains a mystique that most other forms of engineering struggle to emulate, and that none have managed to maintain quite so well. But do not be blinded by the glamor, this is no witchcraft; space technology offers unique opportunities in science, technology and commerce.

1.1 Historical Perspective

On October 4, 1957 the Space Age began with the launch of Sputnik-1 (Russian: PS-1, Спутник-1, or *Elementary Satellite-1*) into an elliptical, low-Earth orbit by the Union of Soviet Socialist Republics (СССР or USSR), and simultaneously a Cold War crisis was initiated within America and her allies. To illustrate the impact of Sputnik-1 in modern terms, the American labor union leader Walter Reuther decried it as a "bloodless Pearl Harbor", invoking the same imagery as used in the aftermath of major modern-day events such as the series of coordinated suicide attacks within the United States of America (US) on September 11, 2001. The Sputnik Crisis within the US heralded the political imperative that space technological superiority would hold, and which would materially drive the development of the technology for both good and bad, especially with regard to human space flight.

The emergence, evolution and development of space technology, like most technologies, can be described by borrowing from Diffusion of Innovations theory and using a logistics function, seen in Fig. 1.1. Starting with the early pioneers, progressing through to the launch of Sputnik-1, and beyond, this section will provide a historical perspective for the later technology chapters of the handbook. However, due to the above-noted political imperative that has been,

M. Macdonald (✉)

Advanced Space Concepts Laboratory, Strathclyde Space Institute,
University of Strathclyde, Glasgow, Scotland
e-mail: malcolm.macdonald.102@strath.ac.uk

P. Norris

Space Defence and National Security, CGI, Leatherhead, UK

D. B. Spencer

Department of Aerospace Engineering, The Pennsylvania State
University, University Park, USA

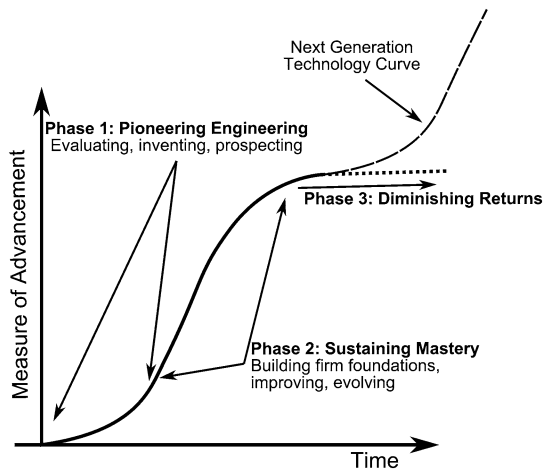


Fig. 1.1 The logistic function, taken from diffusion of innovations theory and used to describe a generic technology development

and still is, placed on human space flight and the lack of societal acceptance of the associated risk, human and robotic space flight will, in this section of the handbook alone, be considered separately.

1.1.1 Pre-Space Age (pre-1957)

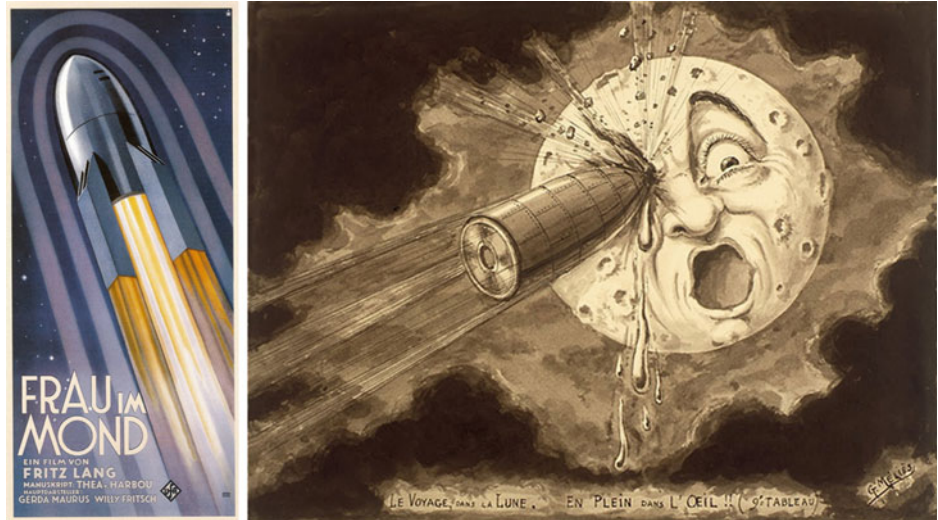
It was not until shortly after the time of Jules Verne's novel *From Earth to the Moon* that space travel became a seriously considered engineering endeavor through the works of Konstantin Tsiolkovsky (1857–1935), a school teacher who would become the father of cosmonautics. Indeed, it is remarkable to note that Tsiolkovsky credited Verne's novel with planting the “*seeds of curiosity*” which led to his calculations of the motion of a rocket. Tsiolkovsky documented many vital features of space travel, detailing these for the first time, and laying the foundations to inspire generations of engineers that followed him. Amongst his work, Tsiolkovsky derived the modern form of the ‘Ideal Rocket Equation’, see [Chap. 4](#), in 1903 [1], prior to the first flight of the Wright brothers, and determined for the first time the velocity required to reach Earth orbit. It should be noted however that rocketry research within western Europe was by this time well established, having been induced by, amongst others, the 1810 Academy of Copenhagen's prize question to calculate the curve described by a rocket when projected in any oblique direction *in vacuo*. Induced by this 1810 prize, mathematicians such as William Moore (fl. c. 1806–1823), the sixth mathematical master at the Royal Military Academy, Woolwich, England considered the motion of rockets in media other than air. The justification of this early rocketry research was principally to advance theories in Naval Gunnery. However, a form of Tsiolkovsky's rocket equation can be found in Moore's 1813 treatise when considering the “*motion*

of rockets in a non-resisting medium” [2], perhaps representing the earliest example of this kind of equation. It is also of note that the Belgian artillery Major-General Casimir Erasme Coquilhat (1811–1890) published an article in 1873 titled *Trajectoires des fusées volantes dans le vide* (*Trajectory of flying rockets in vacuum*) in the *Mémoires de la Société Royale des Sciences de Liège* [3]. Coquilhat's work, like Moore's before, was focused on the use of rockets in war, but, like Moore, he presents the rocket equation before Tsiolkovsky's 1903 publication.

Tsiolkovsky also recognized that multi-stage rockets, when fueled by something akin to liquid oxygen, would be best to achieve this required velocity change to reach Earth orbit. This seminal work was published in 1903 in the Russian journal *Научное обозрение*, or *Scientific Review*, in an article titled *Investigation of outer space rocket appliances* [1]. However, rather unfortunately the journal edition also contained a politically revolutionary article that led to the journal's confiscation by the Tsarist authorities, meaning that Tsiolkovsky's work went virtually unnoticed until after the Bolshevik Revolution; his support of which it can be surmised was not harmed by this earlier incident. Amongst Tsiolkovsky's other work, he designed rockets with steering thrusters, space stations, airlocks for exiting a spaceship into the vacuum of space, and closed-cycle biological systems to provide food and oxygen for space colonies. After the October Revolution of 1917, and the creation of the USSR, Tsiolkovsky's accomplishments were formally recognized and in 1921 he received a lifetime pension from the state that allowed him to retire from teaching and devote himself to his studies. It was sometime after this that his work became more widely known in the West, by which time engineers such as Hermann Oberth (1894–1989), in Germany, and Robert Goddard (1882–1945), in the United States (US), had independently derived many of his key contributions. However, within the Soviet Union Tsiolkovsky was a powerful influence on, amongst others, the luminary engineers Sergei Korolev, also transliterated as Sergey Korolyov (1907–1966) and Valentin Glushko (1908–1989).

The emergence, evolution and development of space technology through the pioneering engineering phase of [Fig. 1.1](#), which commenced with the works of Tsiolkovsky, was continued by Robert Goddard, who is credited with designing and building the world's first liquid-fueled rocket, which he successfully launched on March 16, 1926. As early as 1920 Goddard had proposed the concept of sending a small rocket to the Moon; however this concept was widely ridiculed in the mainstream American press. The *New York Times* wrongly stated in an editorial that Goddard's understanding of Newton's laws of motion was fundamentally flawed due to “*the need to have something better than a vacuum against which to react*”. It took the *New York Times*

Fig. 1.2 *Frau im Mond* movie poster (left) and a pre-visualization painting of the iconic shot from *Le Voyage dans la lune* of the rocket hitting the eye of the Moon (right)



forty-nine years to issue a correction to this statement and an apology; published the day after the Apollo-11 launch. And of Goddard's first liquid-fueled rocket flight? It traveled 56 m in 2.5 s, reaching a maximum height of just over 12 m.

In parallel to Goddard's efforts, Hermann Oberth, again inspired by the writings of Jules Verne, was investigating rocketry and space technology. Oberth discussed almost every phase of rocket travel, including the abnormal effects of pressure on the human body and published many popular books which were important not only for the new ideas within them, but for the inspiration they gave. Indeed Oberth also worked as a technical consultant on the first, so-called, hard science fiction film to have scenes set in outer space, *Frau im Mond*, released in 1929, directed and produced by the film pioneer Friedrich Christian Anton 'Fritz' Lang (1890–1976) and based on the novel *Die Frau im Mond* written in 1928 by Thea Gabriele von Harbou (1888–1954) (Fig. 1.2). This film was key to popularizing the ideas of rocketry and space exploration. However, it should be noted that the first ever science fiction film, *Le Voyage dans la lune*, directed and produced by Georges Méliès (1861–1938, full name Marie-Georges-Jean Méliès), a 1902 French black-and-white 14 min (projected at the then standard 16 frames per second) silent movie, also featured lunar exploration. *Le Voyage dans la lune* was loosely based on two popular novels of its time, *From the Earth to the Moon* by Jules Verne and *The First Men in the Moon* by H. G. Wells.

However, the emerging popular culture interest in rocketry and space travel, along with a small pioneering group of engineers, was not yet reflected within the establishment. This began to change in the early 1930s within Germany. As with so many technical innovations, and as an unforeseen consequence of previous attempts to avoid future wars in Europe, it was the military who would drive the technology forward. Under the terms of the Treaty of Versailles, the Weimar Republic, officially the Deutsches Reich, colloquially known

simply as Germany in the English speaking World, was forbidden from having an air force or advanced artillery. As a result, the German military began to search for new weapons that would not violate the terms of this treaty. In August 1932, approximately six months prior to the foundation of the Third Reich, Wernher von Braun (1912–1977) was recruited by the German Army to aid in developing liquid-fueled rockets, essentially as a form of long-range artillery. It is of note that while von Braun made the decision to accept a research grant from the German military, and hence to be associated with the National Socialist German Workers Party (NSDAP, or Nazi party), other leading compatriots within the German amateur rocketry society, the Verein für Raumschiffahrt (VfR), such as Walter Hohmann (1880–1945), whom the Hohmann Transfer is named after, see Chap. 4, distanced themselves from rocketry, wishing not to make this association. Von Braun's decision to help the Army was, however, consistent with his family's strong tradition of undertaking public service roles¹ and must correctly be viewed within the context of the day. By late 1934, von Braun had developed the Aggregat-2, or A2, rocket and by the end of 1937 the much larger A3, planned as a prototype of the A4, had been launched. However, the A3 had several difficulties, lacking the endurance, control and aerodynamic performance that would be required for the supersonic A4. As a result, the A4 was postponed and work progressed to the smaller A5 in order to resolve the issues encountered with the A3. It is of interest to note that until 1939 and the outbreak of World War II, von Braun was in occasional contact with Goddard, and indeed, many of Goddard's concepts and ideas found their way into the Aggregat series of rockets. As World War II progressed, Adolf Hitler (1889–1945), the then Chancellor of Germany (as Führer und

¹ The definitive English language biography of Von Braun is Michael J Neufeld's *Von Braun—dreamer of space, engineer of war*.

Reichskanzler), warmed to the rocket program and in December 1942 authorized the development of the A4 rocket as a weapon. Re-designated the V-2, the first production missiles were complete in early 1944 and at least 5,000 V-2s were produced by the end of the war. The Nordhausen (Mittelwerk) V-2 factory, along with other V-2 production sites, utilized slave labor, in this case from the nearby Mittelbau-Dora concentration camps. It is believed that as many as 20,000 prisoners died while working at the Nordhausen plant, with as many as 9,000 of these dying from exhaustion. In fact, it is widely held that more people died building the V-2 rockets than were killed by it as a weapon. The V-2 rocket was the world's first long-range military-ballistic missile and the first known human artifact to achieve suborbital space flight. The V-2 rocket is also the foundation on which most modern rockets are based, including those of the United States, the Soviet Union, and Europe's Ariane family.

In the final weeks and months of World War II the Allies each became increasingly keen to capture for themselves as much of the knowledge acquired by the V-2 program as possible, and to deny German engineering knowledge and expertise to the other allied nations. However, von Braun made the decision to surrender to the Americans rather than, as he saw the other options, be captured by the Soviets, or shot by the German command. Despite his status as a *Schutzstaffel* (SS) officer, rank *Sturmbannführer*, and direct links to slave labor, von Braun and his team were 'bleached' of their Nazism by US forces and granted security clearance to work in the US. The UK and Soviets also recovered a number of V-2 rockets, with the UK launching three V-2's from Northern Germany however most of the senior engineers involved had by that stage already agreed to move to the US (Fig. 1.3).

Subsequent to World War II it was assumed within the West, and especially in the United States, that the US had gained a significant technological advantage over the Soviets by acquiring von Braun and his team. As a result of von Braun's surrender to the Americans the post-war development of rocketry within the now Cold War enemies was notably different and it could be argued would, in-effect, be to the long-term advantage of the Soviets with regard to rocket technology, skill-base and manufacturing. The Americans launched several V-2 rockets in the post-war years, notching up several firsts, including, the first scientific experiment in space, a cosmic radiation experiment in May 1946, and the first image of Earth taken from space (altitude 105 km) in October of the same year on-board a V-2 rocket launched from White Sands Missile Range; seen in Fig. 1.4. A collage of images from a similar V-2 launch, this time on July 26, 1948 is shown in Fig. 1.5; the area shown in Fig. 1.5 is approximately 200 million hectares, with a distance to the horizon of over 1,000 km.

In 1947, the Soviets launched the R-1 rocket, which although only a direct copy of the V-2 rocket, based on the



Fig. 1.3 Wernher von Braun with a model of the V-2 rocket. *Image* NASA



Fig. 1.4 First ever photo of Earth taken from space in May 1946. *Image* White Sands Missile Range/Applied Physics Laboratory

few V-2s and associated staff the Soviets had recovered, had been manufactured by Soviet industry, and was not simply a re-assembled V-2. This learning process, together with the brilliance of Sergei Korolev gave the Soviets critical experience of the process required to build a rocket, enabling them to develop future rockets at a more rapid pace than the Americans. Immediately following the R-1 launch the Soviets began work on the R-2 and R-5, based on extensions of the V-2 technology. By 1953 the Soviets, under the design lead of Korolev, often referred to only as 'Chief Designer' by the Politburo, had begun work on the R-7 rocket, a significant step beyond the German heritage of its predecessors and which later became the basis for the Soyuz launcher. The first testing of the R-7 was conducted in May 1957 from what would soon become Baikonur Cosmodrome, located about



Fig. 1.5 An image collage from a V-2 launch on July 26, 1948 from White Sands Missile Range at 100 km altitude. *Image* White Sands Missile Range/Applied Physics Laboratory

200 km east of the Aral Sea in present day Kazakhstan, and a modified version was subsequently developed to launch Sputnik-1. Nikita Khrushchev, the then leader of the USSR, had wanted to launch Sputnik-1 on September 17, 1957 to commemorate the one-hundredth anniversary of the birth of Tsiolkovsky, but technical problems pushed this launch date, and the start of the Space Age, to October 4, 1957.

1.1.2 Space Age (post-1957)

By the start of the Space Age the pioneering engineering phase of space technology development, as shown in Fig. 1.1, was approaching its end, and phase 2, sustaining mastery, was on the horizon. Sputnik-1 was followed only 30 days later by Sputnik-2. By the time the first American spacecraft, Explorer-1, was launched from Cape Canaveral Air Force Station Launch Complex 26 in Florida on 1 February (UTC)/31 January 1958 at the launch site, Sputnik-1 had, 26 days earlier reentered the Earth's atmosphere at the end of its mission. Just as Sputnik-1 marked the start of the Space Age, Explorer-1 marked the start of the Space Race.

During these final years of space technology pioneering engineering an innumerable amount of technical firsts were accomplished. Amongst these highlights include the first communications spacecraft by Project SCORE (Signal Communications Orbit Relay Equipment), in December 1958; the first spacecraft to visit another celestial body in 1959, when Luna-2 impacted the lunar surface in Mare Imbrium, near the craters Aristillus, Archimedes, and Autolycus; the first weather satellite, TIROS-1, in April 1960; and in June of the same year the first reconnaissance, or spy, satellite was launched by the Americans, the Galactic Radiation and Background-1, or GRAB-1, shown in Fig. 1.6. Note that Galactic Radiation and Background was the covername for electronic intelligence (ELINT) Project Dyno, operated by the U.S. Naval Research Laboratory (NRL). The first reconnaissance satellite has been followed by a huge number of similar spacecraft and today, in addition to the classified militarily operated reconnaissance spacecraft, society is accustomed to tools such as Google Earth providing ready access to space-based Earth observation images. Space-based reconnaissance

is not however simply about taking images of the Earth, including as it does signals intelligence-gathering, as was the case with GRAB-1, which can include communications as well as technical and geolocation intelligence; see Sect. 1.2.4.

In 1962, Telstar-1 became the first active, direct-relay communications spacecraft. An indication of the social impact of space technology around this time, and especially the Telstar-1 spacecraft, can be gleaned from the official football of the 1970 FIFA World Cup in Mexico, the Adidas Telstar, named after the spacecraft; see Fig. 1.7. The Telstar football, now considered a design classic, was painted with black and white panels to make it more visible on black-and-white television, making it look like the Telstar-1 spacecraft, which was roughly spherical and dotted with solar panels in a similar fashion.

By the early 1960s space technology development had entered Phase 2, sustaining mastery, with the first geosynchronous spacecraft, see Sect. 4.4.3, Syncom-2 in 1963, and the first geostationary spacecraft, see Sect. 4.4.3, Syncom-3 in 1964, being closely followed by Intelsat-1, Early Bird, as the first commercial communications spacecraft in geostationary orbit in April 1965. In the same month, the first Molniya (Молния), see Sect. 4.4.6, spacecraft was also launched.

Over the next twenty years spacecraft communications technology would revolutionize society and television, enabling 24 hour news, live sport coverage and instantaneous global communication. The first national satellite TV network was established in 1967 in the Soviet Union, using the Molniya orbit, which had been pioneered only two and half years before. By 1976 the Soviets had developed the first operational Direct-to-Home TV communications spacecraft, Ekran (Russian: Экран, *Screen*), providing one TV channel and two radio channels, in the UHF range, broadcasting direct to homes in northern Siberia. Throughout this period, several spacecraft had been to visit both Venus and Mars, including landers on both planets.

By the early 1980s, many spacecraft had been launched into Earth orbit and systems such as the Tracking and Data Relay Satellite System (TDRSS) had spacecraft providing services to other spacecraft, in this case a communications service. It was also during this decade that the Space Shuttle became operational and throughout the 1980s the sustaining mastery phase of space technology's development was

Fig. 1.6 The GRAB-1 team at Cape Canaveral for a spin test atop Transit-2A (*left*) and a display model of a GRAB satellite at the National Cryptologic Museum, Washington, D.C. *Image* Naval Research Laboratory



Fig. 1.7 Illustration of Telstar-1 and the 1970 Adidas Telstar football. *Image* AT&T (*left*) and Adidas (*right*)



clearly continuing at pace. By the close of the decade the flyby of Voyager-2 at Neptune meant that humans had visited every planet in the solar system.²

The commercialization of space began to truly gather pace in the 1990s, with the development of the Ariane 4 launch vehicle, Fig. 1.8, which captured over 50 % of the commercial launch market, and the associated emergence of a mass domestic market for Direct-to-Home TV. The 1990s also saw the launch of the era-defining Hubble Space Telescope (HST), along with subsequent servicing missions of this and other spacecraft by the Shuttle Orbiter. In 1994, another ubiquitous feature of our modern lifestyles became operational; the Global Positioning System, GPS, a space-based global navigation satellite system, GNSS, operated by the US Department of Defense. Within 15 years of the GPS system becoming operational, several studies within the US estimated that 6–7 % of the Gross Domestic Product, GDP,



Fig. 1.8 The 1st Ariane 4 launch in June 1988. *Image* ESA

² Note that Pluto was still categorized as a planet at this time, but is now classified as a dwarf planet.

of developed countries, and up to perhaps 10 % in the US and EU, could be classed as dependent on GPS [4].

As the Space Age approached the half-century it was clear, looking back, that the sustaining mastery phase of development was approaching its completion, with space technology now an integral facet of daily life, often taken for granted and the importance of which was, and remains, often overlooked. Indeed, by the time the Space Age actually reached its half-century space technology, once the preserve of elite government research groups was accessible to student engineers and scientists through platforms such as CubeSats, the first of which were launched in 2003. Indeed, it could even be argued that missions such as NASA's MESSENGER (MErcury Surface, Space ENvironment, GEochemistry and Ranging), and the ESA/JAXA (Japan Aerospace Exploration Agency) BepiColombo mission, both to Mercury, or the Mars rovers Spirit and Opportunity, together with recent successful developments in space technology by China, India, Brazil and others, demonstrate the end of the sustaining mastery phase of space (robotic) technology development, and the beginning of the diminishing returns phase. That is, today, should humanity choose to do so, a robotic spacecraft could be placed into orbit about, or onto the surface of, almost any planet in our solar system. Furthermore, inserting a spacecraft into Earth orbit, while not mundane, is considered by many to be a routine engineering activity, with commercially acceptable levels of risk.

1.1.3 The Future of Robotic Space Technology

Future forecasting is always unwise. However, as an engineer the possible and perhaps even the probable can at least be identified, hopefully without undue bias to preferable over probable futures. History is of course, littered with over, and under, optimistic predictions, and engineers are by their very nature largely a technologically optimistic band. Yet in 1961, the same year as the first commercial communications satellite entered service in geostationary orbit, T.A.M. Craven, the Commissioner of the US governmental agency the Federal Communications Commission, and an engineer by trade, stated "*There is practically no chance communications space satellites will be used to provide better telephone, telegraph, television, or radio service inside the United States.*"

It would seem that the future of space technology could take two possible paths. The first would be the *business as usual* route, taking the technology along the currently established technology trend, shown in Fig. 1.1, into the diminishing returns phase. The second possible path is that a new innovation leads to the establishment of a next generation technology curve, as shown in Fig. 1.1. While this may

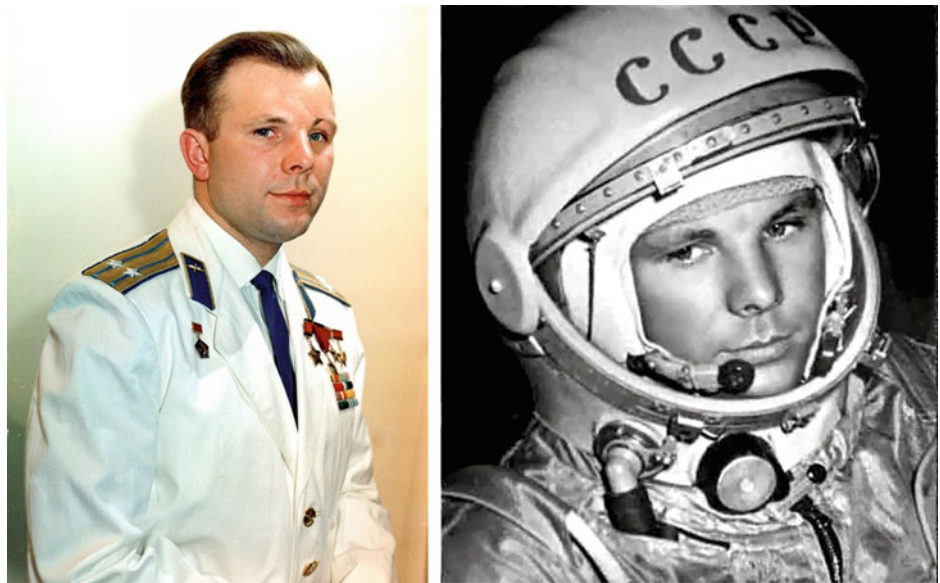
seem like hedging, it is not a matter of whether the second path will happen, but rather when, what requirement the new innovation will fulfill, and hence what the key system level innovation will be.

At present, spacecraft provide a vantage point to acquire and re-direct information; Earth observation, including military reconnaissance, telecommunications and science are all information services. Attempts to exploit microgravity and hard vacuum for scientific or commercial ends have had only limited success. It can thus be theorized that the next generation of space technology will alter this in some way. To this end, and borrowing the language of the Internet, the current generation of space technology can be termed Space 1.0, having created a basic in-orbit infrastructure and market economy, much in the same way as the Web 1.0; Space 2.0 would then open up space access beyond the current national and international participants, exploiting and building on an established infrastructure. Some but not all consider that Space 2.0 could therefore facilitate sharing, interoperability, user-centered design and collaboration. It could also thus be cheaper, with shorter program durations enabled through a reliance on in-orbit infrastructure, services and products. Future low-cost, responsive space missions could use in-orbit infrastructure rather than build a complete new spacecraft. As such, the terrestrial concept of service utilities could expand into space, with spacecraft becoming interdependent, interoperable and increasingly specialized. The European Data Relay Satellite (EDRS) program is an early example of how this may work on a commercial basis. With such a move, the in-orbit infrastructure could move towards a state of continuous evolution and maintenance, rather than the current generational step-changes in capability. This type of service-level evolution is already evident in the GPS system. However, presuming that the next generation space technology curve were to begin today, then it must be presumed that the pioneering engineering phase of this curve would extend for the next 20–40 years, and that, of course, assumes that such innovations fulfill a suitable market requirement.

1.1.4 Human Space Flight

The great romanticism of human space flight has sparked many imaginations, Jules Verne's inspirational novel *From the Earth to the Moon* was not about a mechanoid, or a robotic Moon mission, but three human travelers. Throughout the works of Tsiolkovsky, Oberth, and his involvement in *Frau im Mond*, and onto von Braun, each had a true fascination with human space flight. Similarly, Korolev's Sputnik program was, from the very start, about human space flight and not commercial exploitation, or scientific exploration of space.

Fig. 1.9 Yuri Gagarin. *Image*
The Russian State Archive of
Scientific and Technical
Documentation, RGANTD



The first living things intentionally sent into space were fruit flies, on-board a US launched V-2 rocket on February 20, 1947. The rocket reached an altitude of 109 km in just over 3 min. During the descent the capsule was ejected, the parachutes successfully deployed and the fruit flies recovered alive. Several animals subsequently flew on suborbital arcs, including Albert II, a Rhesus Monkey, who became the first primate in space on June 14, 1949. Sputnik-2, the second ever spacecraft to orbit the Earth carried the first animal to orbit the Earth, a dog named Laika; a female part-Samoyed terrier. Laika however survived for only a few hours in orbit instead of the planned ten days due to thermal control system issues and the stress of the experience.

The first test flight of the Soviet Vostok spacecraft was Korabl-Sputnik-1, also known as Sputnik-4 (Sputnik-3 was a robotic spacecraft originally intended to be the first satellite). Despite the conspiracy theories, there is little reason to doubt that this was an automated test flight, with Korabl-Sputnik-2 (Sputnik-5), launched on August 19, 1960, being the first spacecraft to safely return animals sent into orbit. Korabl-Sputnik-2 carried two dogs, Belka and Strelka, 40 mice, and two rats, as well as a television camera that took images of the dogs; it is of note that these images are often mistaken for images from Sputnik-2. It is also of note that one of the dogs suffered seizures during the fourth orbital revolution, as a direct result of this it was decided the first human flight should make no more than three revolutions of the Earth. The first human space flight was subsequently achieved on April 12, 1961, with the launch of the Soviet cosmonaut Yuri Gagarin (1934–1968), who completed a single revolution about the Earth in Vostok-1 prior to reentry (Fig. 1.9).

Considering the development of human space flight, it is apparent that by the time of Gagarin's flight, pioneering

engineering was still very much the order of the day. Indeed it would be the following February before America matched the feat of human orbital space flight with the launch of John Glenn (born 1921). However, in November 1960, John F. Kennedy was elected president of the US, promising superiority over the Soviet Union in both space exploration and missile defense, and warning of a missile gap between the two nations. In fact, the missile gap was a figment of exaggerated intelligence estimates based largely on domestic US politics and inter-service budgetary tensions. However, despite this and Kennedy's own rhetoric it was actually the flight of Gagarin that prompted Kennedy's support for the Apollo program which had been conceived early in 1960 during the administration of Dwight D. Eisenhower (1890–1969). And, six weeks after Gagarin's flight, 20 days after Alan Shepard had become the first American in space on a sub-orbital flight and nine months before the first American orbited the Earth, Kennedy committed the US to "*land a man on the Moon and return him safely to the Earth*", before the end of the decade. Such a goal was hugely ambitious, requiring the US to commit the largest amount of resources by any nation in peacetime towards a single goal, reported to Congress as 25.4 billion USD in 1973, and in 2009, by NASA, as 170 billion USD in 2005 dollars; of course, this was no ordinary peacetime (Fig. 1.10).

Returning to the logistic curve of Fig. 1.1 it is apparent that the Apollo program is too early. The political imperative placed on human space flight means a requirement emerges to develop sustained mastery of a narrow specific application of political value, prior to completing the necessary pioneering engineering. As such, despite the clear and obvious technical achievements of the Apollo program, along with its cultural impact, its long-term value is questionable. Indeed, modern proposals from within the US to

Fig. 1.10 A 1964 portrait of Neil Armstrong (*left*) and Buzz Aldrin on the porch of the Apollo-11 lunar module, Eagle (*right*). *Image NASA*



Fig. 1.11 Mir space station and insignia. *Image NASA*



return to the Moon, such as NASA's canceled Constellation program, highlight the limited long-term value provided by the Apollo program.

Throughout the Apollo program, the Soviet Union, in addition to attempting its own, abortive human lunar program, continued to develop its Earth orbiting human space flight program. The Soviets built on early successes, working towards the development of the Soyuz capsule and launcher, which first flew a cosmonaut into space in 1967 killing him and which continues today to operate as a derivative of this original vehicle. The first Soyuz flight, Soyuz-1, had an ambitious mission plan, including a rendezvous with Soyuz-2 however due to a range of technical issues this did not happen until Soyuz 4 and 5 flights. On return to Earth, cosmonaut Colonel Vladimir Komarov

(1927–1967) died when the spacecraft crashed, becoming the first in-flight fatality in the history of space flight. On April 19, 1971 the Soviet Union launched Salyut-1 (DOS-1), the first space station and the basis on which the Mir space station was developed and operated from 1986–2001 (Fig. 1.11). Following the cancellation of the Apollo program, the US developed Skylab, the first American space station, which hosted three crewed missions in 1973–1974, and flew until 1979 bridging the gap from Saturn V and the Apollo program to the Space Shuttle program. The Shuttle Orbiter flew from 1981–2011 and although once again a significant technical accomplishment, by the time it was decommissioned the Americans had once again failed to sustain the technology in anything close to that of the by now mature, established and work-horse Soyuz.

Fig. 1.12 Europe's Columbus Laboratory attached to the International Space Station.
Image NASA



Prior to the retirement of the Shuttle Orbiter, it was largely responsible for the establishment of the International Space Station, ISS; the most expensive international collaborative scientific endeavor ever undertaken by humanity, estimated to have a similar total cost to the Apollo program in real terms. On orbit construction began in 1998 and was almost completed by the final Shuttle Orbiter missions over a decade later (one large and a few small Russian elements of the ISS were still on the ground when the Shuttle fleet was retired) (Fig. 1.12).

1.1.5 The Future of Human Space Flight Technology

The technology development of human space flight and its position on the logistic curve of Fig. 1.1 is difficult to accurately quantify. The Soviet Union, and now Russia, has certainly developed a sustained mastery of low-Earth orbital flight with the Soyuz, which has effectively been operational for well over 40 years. Meanwhile the US has failed to build such a distinguished heritage of human space flight technology and the safety record of the Shuttle Orbiter was disappointingly poor. However, there should be little doubt that the US has a significant level of mastery of this technology, if not the politics associated with it. Beyond the two traditional powers, China was the third nation to develop an independent human space flight capability in 2003, using the Soyuz spacecraft's heritage. However, they, together with the other emerging powers, cannot yet be classed as anything other than exactly that, and it will take a significant period of time for mastery of the technology to be established.

It is therefore likely that the immediate future destination of human orbital space flight is the ISS and low-Earth orbit. Providing humanity with a chance to develop a globally sustained mastery of the required technologies for humans to live in space for prolonged periods, while understanding the physiological and psychological effects of such expeditions,

with some important exceptions such as exploring how artificial gravity can alleviate those physiological effects. Only after humanity has mastered the technology of low-Earth orbit can it consider moving beyond this. However, in a memorandum from Robert S. McNamara, the then US Defense Secretary, and James E. Webb, the then NASA administrator, to Vice President Lyndon B. Johnson, dated May 8, 1961, a mere 17 days before Kennedy announced the Apollo program to the US Congress, it was argued that achievements in space “symbolize the technological power and organizing capacity of a nation” [5]. Hence, it must be recognized that the romanticism of human space flight, along with the perceived national prestige, have held, and will continue to hold, the interest of politicians, securing funding for engineering, science and mathematics research and educational programs. Hence, even if the concept that human space flight “needs somewhere to go” is from a bygone Cold War era of a race for political ideological superiority, the principle of human space flight beyond the Earth's gravity well will always have significant value beyond the feat itself.

1.2 Applications of Space Technology

The US National Space Policy [6] distinguishes three types of space activity, commercial, civil, and national security. That categorization is followed in this Handbook with the exception that ‘science’ applications are removed from the ‘civil’ category and treated as a separate, fourth, category.

1.2.1 Civil

Civil space flight covers publicly funded space programs that do not inherently address scientific or security objectives. The wide range of applications in this category is illustrated by the following examples, each of which will be described in more detail

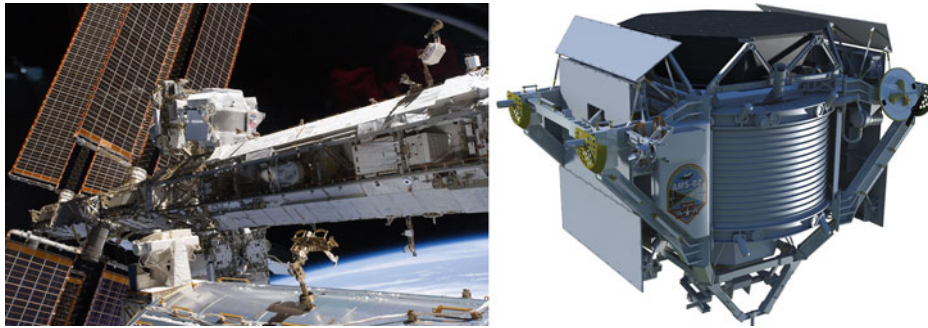


Fig. 1.13 Starboard truss of the International Space Station on 20 May 2011, the Alpha Magnetic Spectrometer (AMS-02) is visible at center left (left) and a computer generated image of the AMS-02 (right). Image NASA

- Human space flight
- Operational meteorology
- Search and rescue
- Technology demonstration
- Education.

1.2.1.1 Human Space Flight

Since Yuri Gagarin orbited the Earth on April 12, 1961 more than 500 people have been in space. The US, Russia, Japan, Europe and Canada undertake the vast majority of current human space flight activities and, as discussed in Sect. 1.1.5, currently focus most of their efforts on a collaborative endeavor centered on the International Space Station.

Having taken more than a decade to assemble, the ISS has now entered its operational phase. It provides facilities for the following main areas of research [7]:

- *Biology and Biotechnology*—In microgravity, controls on the directionality and geometry of cell and tissue growth can be dramatically different to those on Earth. Various experiments have used the culture of cells, tissues and small organisms in orbit as a tool to increase our understanding of biological processes in microgravity.
- *Physical Sciences*—The space station provides the only place to study long-term physical effects in the absence of gravity. This unique microgravity environment allows different physical properties to dominate systems, and these have been harnessed for a wide variety of physical sciences.
- *Multipurpose*—From freezers and incubators, to glove boxes and complete racks, standard multi-purpose facilities support a wide range of research on-board the space station.
- *Earth and Space Science*—The presence of the space station in low-Earth orbit provides a vantage point for collecting Earth and space science data. The largest space science facility on the station is the AMS-02 (see below).
- *Human Research*—The space station is being used to study the risks to human health that are inherent in space exploration, especially as concerns long-duration residence in a

microgravity environment. This research is also relevant to the study of some Earth-bound conditions such as the effects of long bed rest, osteoporosis and diseases such as bone marrow depletion.

- *Technology*—Studies on the space station can test a variety of technologies, systems, and materials that will be needed for future long-duration exploration missions.
- *Educational Activities*—Station educational activities have had a positive impact on thousands of students by involving them in station research, and by using the station to teach them the science and engineering that are behind space exploration. Long-term benefits include inspiring students to excel in mathematics and science.

The Alpha Magnetic Spectrometer (AMS-02) was delivered to the International Space Station on Space Shuttle Endeavour's last mission (flight STS-134) in 2011. AMS is a particle physics detector, designed to search for various types of unusual matter by measuring cosmic rays. AMS-02 weighs just less than 7 metric tons, uses 2 kW of electrical power, and has a design operational lifespan of a decade. The design and development was led by MIT Professor and Physics Nobel Laureate Samuel Ting (1936–present). AMS-02 is the most significant attempt to-date to measure high-energy cosmic rays directly—instead of via the secondary or tertiary products of their collisions with Earth's atmosphere (Fig. 1.13).

China has a separate and active human space flight program, while India has begun a similar program and Iran has stated that it plans to follow suit. China's human space flight program began in 2003 with the flight of Yáng Lìwěi (born 1965) in Shenzou-5 (Divine Ship) for 21 h. Of note, official English-language texts issued by the government of the People's Republic of China use the term 'astronaut' to describe professional space travelers from China. However, the term 'taikonaut', a hybridization of the Chinese term 'taikong' (space) and the Greek 'naut' (sailor), is often used by English-language news media organizations and likely has its origins in the term 'tàikōng rén', 'spaceman', used in Taiwan (formally the Republic of China) and Hong Kong

Fig. 1.14 Schematic of the world's operational weather satellites in early 2011. *Image* EUMETSAT

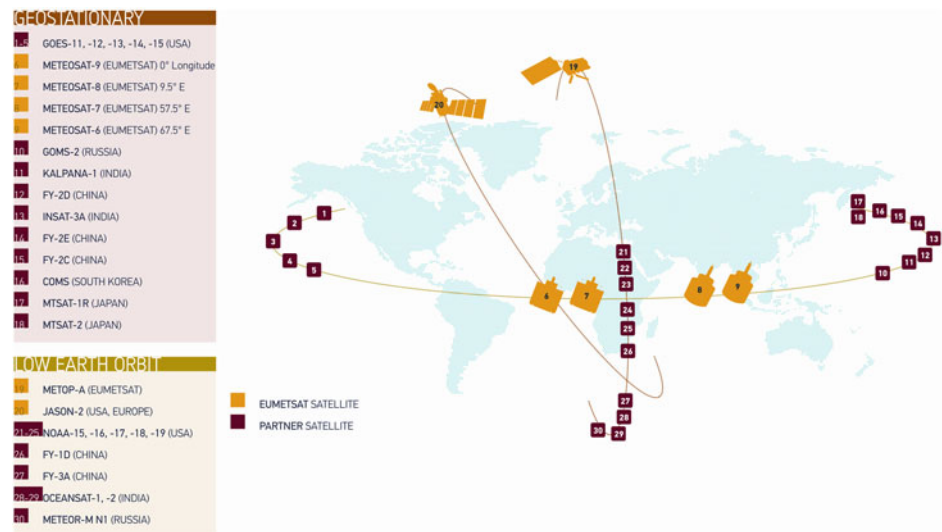
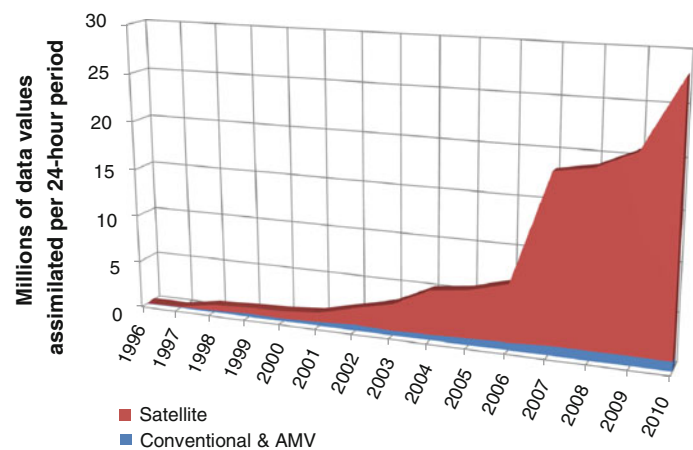


Fig. 1.15 Conventional and satellite data assimilated at ECMWF. *Data source* ECMWF



(a special administrative region of the People's Republic of China) to mean professional space travelers. Meanwhile in Chinese the terms 'yǔ háng yuán' (Chinese: 宇航员), and 'háng tiān yuán' (Chinese: 航天员), both meaning 'astronaut', have been used historically to describe professional space travelers and recently some influential English language media organizations, such as the BBC, have begun to use the transliteration 'yuhangyuan' rather than the term 'astronaut'. In a steadily paced program, China gradually increased the scale of its human space flight missions from its first flight in 2003 through to demonstrating extra-vehicular activity, EVA, capabilities during Shenzhou-7 in 2008, and the autonomous docking of spacecraft, initially with the crewless Shenzhou-8 with the much larger Tiangong-1 module in 2011, and the crewed Shenzhou-9 in 2012.

1.2.1.2 Operational Meteorology

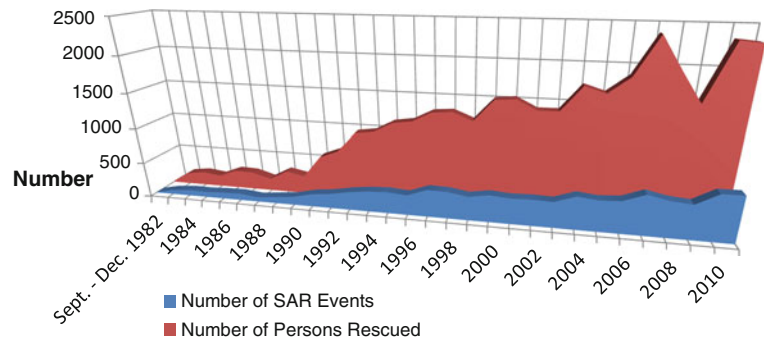
As already mentioned, the first weather satellite, TIROS-1, was launched in 1960 by the US, demonstrating the value of space-based observations to assist weather forecasters.

Today, a fleet of satellites owned by seven countries and regions (Europe is counted as a single region in this tally) monitor the Earth's weather on a routine basis from two basic orbital positions—near-polar orbits of about 900 km altitude and geosynchronous orbits 36,000 km above the Earth, see Fig. 1.14 [8].

The dependence of medium term (five to 30 days) weather forecasts on satellite data is illustrated in Fig. 1.15, which shows the increasing dominance of satellite data in the forecasts of the European Centre for Medium-range Weather Forecasts (ECMWF). The dominance is even larger than at first glance because the 'conventional' graph includes Atmospheric Motion Vectors (AMV—in other words, wind vectors) calculated by tracking cloud movements in satellite images. ECMWF found that the conventional data alone gave such poor results that it was necessary to add the AMV data in order to get a basic forecast against which the benefit of adding other satellite data could be assessed.

The satellite data considered essential for modern weather forecasting include (items in brackets are slightly less essential)

Fig. 1.16 Number of SAR events and persons rescued with the assistance of Cospas-Sarsat alert data (121.5 and 406 MHz) between January 1994 and December 2011



- (At least six) geostationary satellites each with
 - Visible/infrared multi-spectral imager
 - (Infrared hyper-spectral sounder)
 - (Lightning imager)
- Polar orbiting sun-synchronous satellites (in 3 orbital planes) each with:
 - Visible/infrared multi-spectral imager
 - Microwave sounder
 - Infrared (hyper-spectral) sounder.

Other sensors of demonstrated interest to weather forecasters include microwave imagers, scatterometers, radio occultation (constellation), altimeter (constellation), infrared imager (for sea surface temperature), more and advanced versions of the visible/infrared imagers (for ocean color and land cover), precipitation radars, broadband visible/infrared radiometers (for radiation budget), instruments to monitor atmospheric constituents, and imaging radar.

Future space-based sensors which may enter operational use include Doppler wind lidar, low frequency microwave radiometer (for salinity and soil moisture), gravimetric sensors (for water), microwave imager/sounder in geostationary orbit (for precipitation), advanced imagers in geostationary orbit and imagers on satellites in other orbits (e.g., elliptical, high inclination) [9].

1.2.1.3 Search and Rescue, SAR

About a dozen polar orbiting satellites carry equipment to detect emergency transmissions from Cospas-Sarsat beacons—this number will increase dramatically when the Galileo navigation system is deployed (see Sect. 1.2.2), ensuring immediate detection of an alert anywhere in the world. The majority of the emergencies are at sea, but some are also from people in distress on land or in aircraft. Each ship, plane or other user purchases the special beacon, which in some circumstances is mandatory. Alerts detected by Cospas-Sarsat are passed to search and rescue authorities in the 41 participating countries. An overview of the SAR events assisted by Cospas-Sarsat between 1994 and 2011 is shown in Fig. 1.16, data from [10]. It is noted that a possible explanation for the reduction in SAR events in 2009 was the

phasing out of the 121.5 MHz alert system in February of that year, at which point the beacon population was estimated at nearly half a million, or about 30 % of all beacons.

1.2.1.4 Technology Demonstration

Many countries test technology in orbit before using it operationally. The items to be tested may be piggy-backed on a satellite dedicated to other purposes or may occupy a special satellite dedicated to the testing of technologies. The justification for placing an object in orbit just to test it is that ground testing cannot replicate the conditions in space accurately. For example, simulating zero-gravity on Earth is only possible for short periods, so any equipment for which zero-gravity may prove problematic needs to be proven in orbit.

In general, a purchaser of an expensive satellite will feel more confident of its success if all of its components and subsystems are of a design that has previously been orbited successfully. Similarly, the credentials of a designer or manufacturer, especially one with little or no record of accomplishment, will be enhanced by successful flight of equipment on a test satellite.

1.2.1.5 Education

Besides acting as an inspiration for students, space contributes directly to education primarily in the form of telecommunications. India has found the use of satellites for communicating with widely dispersed schools so beneficial that the EDUSAT satellite, or GSAT-3, was launched in 2004 and is dedicated to linking schools across the sub-continent—sometimes referred to as tele-education, analogous to tele-medicine in the medical arena. Edusat is part of a long-term commitment of India's space agency, ISRO, to using satellites for education.

1.2.2 Commercial

1.2.2.1 Satellite Communications and Broadcasting

World-wide commercial space activities were valued at 304 billion USD in 2013, an increase of 7 % over the previous year, and comprised of 226 billion USD for commercial

infrastructure, support industries, space products and services, and 78 billion USD for government space budgets [11]. Direct-to-Home television services represent the largest sector of this activity, while other sectors include broadband communications, navigation, remote sensing (Earth observation), space tourism, and microgravity, each of which is outlined in one of the following sections.

The main reasons why satellites are attractive for communications and broadcasting include the following

- Coverage over regions varying in size from a city to a continent via a single space-based hub (transmitter/receiver) where the alternatives require either thousands of Earth-bound hubs, or wires and cables connected to every consumer.
 - A satellite solution therefore is often cheaper and/or faster to deploy and causes less disruption (digging up roads, constructing hubs, etc.).
- Licensing of a single wide-area satellite may be simpler than licensing a multitude of Earth-bound mini-systems to provide the same coverage.
 - Examples abound of satellites being used to establish communications between widely separated countries avoiding the fees, licensing and other commercial and regulatory barriers associated with using terrestrial links.
- Full quality communication with ships, planes and land vehicles far from terrestrial infrastructure is readily provided by satellites where the alternatives such as short wave radio are unreliable and/or of low quality.
 - The rapid provision of temporary links to anywhere in the world is especially attractive to the news and sports media.
 - Communication with other satellites, typically to relay information beyond the line of sight, is primarily a public sector service (principally for military and human space flight missions) but the European Data Relay Satellite (EDRS) program is an attempt to place such a service on a commercial footing.
- International spectrum regulations define certain spectral bands for satellite services which satellites can use either once per region with a single beam or multiple times via spot beams that can be as small as city-sized.
- Geostationary satellites are especially attractive for commercial services since a single satellite can provide communications and broadcasting services.
 - Low-Earth orbit satellites benefit from a much shorter transmission distance but a constellation of several tens of satellites is required to ensure continuous service of all parts of the globe thus requiring close to the full constellation to be in place before 24 hour service can be offered—a much more expensive investment than a single geostationary satellite.

Satellite TV subscriptions are forecast to continue to grow, especially in Asia-Pacific (e.g. India) so that developing countries will represent 70 % of subscriptions by 2020, up from 45 % in 2010 [12]. Satellite broadband is also forecast to grow nearly ten-fold by 2018, especially in North America [13]. Mobile satcom terminal numbers are expected to double by 2020 especially in Machine-to-Machine (M2M, asset tracking/telemetry) applications, and mobile broadband is forecast to grow to about 10 % of mobile terminal numbers by 2020 from zero in 2008—and much more than 10 % in terms of revenue [14]. The provision of broadband via satellite to homes and businesses has given rise to the first satellites with capacity exceeding 100 Gbps. Examples include the 140 Gbps satellite Viasat-1, launched in October 2011, which provides downloads in the range of 8–12 Mbps for users. As of 2011, the bandwidth capacity offered by the world commercial satcom market is split across the spectrum roughly as follows: Ka band is 20 %, Ku band is 50 %, C band is 20 % and L, S and X bands together are 10 % [15].

1.2.2.2 Global Navigation Satellite System

The expensive investment required to deploy a constellation of non-geostationary satellites was mentioned in the previous section. For satellite-based navigation this investment has been made by several governments around the world; the world's main satellite-based navigation systems are listed in Table 1.1, from [16]. Non-military users world-wide are benefitting from the free availability of the Global Navigation Satellite System (GNSS) signal in space, and as previously mentioned up to perhaps 6–10 % of the GDP of developed countries may be classed as dependent on GNSS.

Mass-market GNSS receivers are often embedded in a general purpose device such as a mobile phone. Many others are tailored for use in vehicles and packaged with a general purpose microprocessor containing a digital map, a route optimization algorithm, a map-matching algorithm (adjusts the user's location to be on a road), databases of places of interest, a touch screen, etc. The fusion of satellite navigation with these other mass market devices means that the number of users with suitably equipped devices is in the hundreds of millions, although the number of, for example, phone users who actively use the GNSS functionality is difficult to measure.

The features of a satellite solution for navigation that make it attractive include

- A single system works anywhere in the world (with a view of the sky) thus reducing the investment of the receiver manufacturers.
- The space infrastructure is funded by either the military or high-value commercial contracts, providing a free, but degraded public-service.
- The accuracy is in the meter range, compared to 50 m to 1 km from cell phone-based techniques.

Table 1.1 The world's main satellite navigation systems as of 2013

US: Global Positioning System (GPS)
24 satellites are the minimum required to provide continuous worldwide service. Each satellite circles the Earth in about 12 h at an altitude of 20,000 km. Some transmissions are encrypted and on special frequencies intended for military users
US/Europe/Japan/India: Wide Area Augmentation System (WAAS)/EGNOS/MSAS/GAGAN
The Federal Aviation Authority operates WAAS as an adjunct to GPS, alerting users in the US (including Hawaii and the southern half of Alaska) within 6 s if any GPS satellite is not working properly, and providing users with information to reduce the errors, especially those caused by electrical storms in the ionosphere. Geostationary satellites are used to get the WAAS information to users rapidly and over a wide area. EGNOS, MSAS and GAGAN are the equivalent GPS augmentation systems in Europe, Japan and India
Russia: GLONASS
GLONASS operates on a similar principle to GPS, with satellites about 19,000 km above the Earth orbiting every 11¼ hours. 24 satellites are needed for a continuous world-wide service. Some transmissions are encrypted and on special frequencies intended for military users
Europe: Galileo
Due to enter limited-global service in 2014 and to be complete around 2020, Galileo is similar to GPS, with an eventual 30 satellites 23,000 km above the Earth orbiting every 14¼ hours. Uniquely among global coverage systems, Galileo will operate on a commercial basis with some transmissions encrypted and on special frequencies intended for subscription-based users
China: BeiDou-2 (or Compass)
Began offering location, timing and navigation services to China and surrounding areas in late 2011, targeting global coverage by 2020; uses a mix of geostationary, geosynchronous and GPS-like orbits. As with GPS and GLONASS, some transmissions will be encrypted and on special frequencies intended for military users
India: IRNSS
Seven satellites will be placed in geosynchronous orbit by approximately 2014 to offer GPS-type signals over the South Asia region
Japan: QZSS
Three geosynchronous satellites will provide GPS-type signals plus some messaging services over the Japanese region. The first satellite was launched in 2010
International: Cospas-Sarsat
Twelve satellites currently carry the equipment to pick up signals from special Cospas-Sarsat emergency beacons

- The GPS system is receive-only thus eliminating the need for transmission (transmission is inherently more energy intensive than reception).
 - The Cospas-Sarsat system requires the user to transmit an emergency message.
- The receive-only aspect of GPS means that the privacy of users is protected—in contrast for example to using the mobile phone network to locate a user, which by definition makes that location known to the mobile network.

1.2.2.3 Earth Observation

A space-based sensor for observing the Earth has many attractions in the commercial sector

- Once the satellite is deployed it can take as many images as required each time it views the area of interest—unlike an aircraft-borne sensor which requires a dedicated flight of the aircraft each time.
 - A geostationary satellite can provide continuous viewing of a large part of the globe.
 - The 36,000 km altitude of such a satellite makes imagery of better than about 1 km resolution challenging.
 - Satellites do not require legal authorization to view the Earth thus making it possible to view any country and to view across borders (for example to analyze geological structures or water resources).
- Of course, there are many disadvantages to using a satellite to observe the Earth, including
- The large distance (a typical observing satellite is in orbit at an altitude of 600–800 km).
 - The need for sunlight and cloud free skies
 - Imaging radar sensors offer a solution to this but they are more expensive than optical sensors and provide very different information.
 - Infrared sensors can provide night-time images but with less resolution and contrast than images from similarly priced visible light sensors.
 - The sometimes long delays between revisits to the same area.
 - Satellite constellations such as RapidEye are intended to address this problem.
 - The extra cost of a satellite sensor over that of an aircraft-borne one because for example it has to work in vacuum and zero-gravity without maintenance, to survive launcher vibrations, to be miniaturized and designed to use minimal electrical power, etc.

For these and other reasons the commercial Earth observation market had taken 20 years to reach world-wide sales of about \$100 million at the turn of the century. In the first decade of the 21st century, it has grown ten-fold to more than \$1 billion per annum aided by the following factors

- The US military and intelligence agencies decided to outsource their requirement for imagery of the 50 cm resolution class to industry, offering decade-long multi-billion dollar contracts to two companies, GeoEye and DigitalGlobe.
 - The same US government agencies have also outsourced some of their requirement for 1 m class imaging radar to three non-American satellite companies but over a shorter (3 year) period and with a funding level that is two orders of magnitude smaller.
- Google has made imagery of the whole Earth available to Internet users free of charge, thus familiarizing millions of people with space-based imagery.
 - Google Earth imagery is from a mix of satellite- and aircraft-borne sensors (the latter especially in developed regions).
- The cost of space-based sensors with a useful imaging capability has fallen sharply driven by advances in digital electronics and cameras.
 - About 20 countries now operate their own imaging satellites that provide imagery with better than 5 m resolution and more countries are set to join this community in the coming decade.

Imagery of the Earth is used in a myriad of applications. Early adopters included the farming and mineral exploration sectors, but today the list of users is extremely broad since to some extent every visitor to Google Earth is a user whatever the reason for that visit. In general, Earth imagery is used to make, update or enhance maps—increasingly digital maps. The wide area form of space-based imagery is ideal for mapping. Stereo imagery is available from several satellite operators, easing the identification of map features. Analysis of optical or radar imagery of an area taken from two different angles allows the height of the surface to be mapped, i.e. provides a 3-D map. By 2011, the state-of-the-art world-wide 3-D map was probably the map developed from the optical images of JAXA's ASTER optical sensor on NASA's Terra satellite. That should be eclipsed in 2014 by one based on imaging radar data from the German commercial TerraSAR-X and Tandem-X satellites—it promises to provide a global map with 2 m local accuracy and 12 m granularity.

A new form of observation from space has been demonstrated by the US Gravity Recovery and Climate Experiment (GRACE) satellite. In the first 5 years of the mission, GRACE measured the Earth's gravity field with unprecedented accuracy and detected significant changes in gravity

over certain regions of the Earth. Over northern India and southern California analysis of the change in gravity suggested a reduction in the underground freshwater aquifers in those regions. Over Greenland, the changes indicated a loss of glacial ice that correlated well with estimates based on satellite altimetry and other measurements [17].

1.2.2.4 Space Tourism

In addition to the more than 500 state-sponsored astronauts and cosmonauts who have been in space, a handful of private individuals have paid to do the same—the price tag for a few days in orbit is typically 20 million USD, or more. Such individuals are termed 'space flight participants' to distinguish them, and other special travelers, from the career astronauts who form the crew of such flights.

Suborbital tourism will soon be offered at a price two orders of magnitude less. Virgin Galactic or Space Expedition Corporation (SXC), using the XCOR Aerospace Lynx horizontal-takeoff, horizontal-landing (HTHL), rocket-powered space-plane, are likely to be the first operator of such a service with operational flights due to start, perhaps, in the 2014–2015 period. Virgin Galactic's SpaceShipTwo service builds on technology developed by Scaled Composites Inc. in winning the suborbital rocket X-Prize in 2004—an air-launched rocket plane tops out above 110 km altitude before gliding back to land. The passengers experience about 10 min of weightlessness. Other potential suborbital space tourism operators are waiting in the wings to see how Virgin Galactic's business develops. Such technology and business development could offer a route towards point-to-point suborbital space flight.

1.2.2.5 Microgravity

The lack of gravity in space suggests the possibility of creating ultra-pure materials such as crystals, lenses and semiconductors, or of growing plants with unusual properties. Many of the required facilities exist on the International Space Station, and now that the station is fully assembled the opportunity is there for the commercial sector to grasp. It remains to be seen whether a market in microgravity activities will develop.

1.2.3 Science

In the context of space flight (excluding human space flight), science can be broadly considered as split into looking outwards for astronomy and space science, and looking downwards to the Earth.

1.2.3.1 Astronomy and Space Science

The following are some of the reasons to do astronomy in space despite the inherent high costs involved

- Observe the sky in those parts of the electromagnetic spectrum that are absorbed by the Earth's atmosphere.
 - Ultraviolet, X-ray and gamma ray at the wavelengths shorter than visible light observing high-energy events such as supernovas.
 - Infrared and millimeter wave at longer wavelengths observing cooler and lower energy events, or events that have been red-shifted by the Doppler effect caused by their rapid movement away from Earth (this last being one of the main justifications for the James Webb Space Telescope).
 - Examples of the many satellites benefitting from this lack of atmospheric absorption feature include XMM-Newton (ESA X-ray), Compton (NASA gamma ray), Spitzer (NASA infrared) and Planck (ESA millimeter wavelength).
- Observe in visible light without the blurring and dimming caused by the Earth's atmosphere; the Hubble Space Telescope (NASA/ESA) is the best-known example of a satellite exploiting this feature of space. Note that ground-based telescopes are increasingly able to eliminate atmosphere-induced blurring by detecting and counteracting the movement of the atmosphere—techniques used to counteract include moving the optics, termed adaptive optics, or the detector or the detected image in a computer, the approach taken is similar to the anti-shake feature in a mass market digital camera.
- Observe the full sky simultaneously with a single instrument. A full-sky survey requires at least two ground-based telescopes, one in the northern and the other in the southern hemisphere, making it difficult to compile a rigorous map of the sky. ESA's Hipparcos provided the best reference star catalog to date when it was released, because of its ability to observe the complete sky and this will be improved only when ESA's Gaia full-sky survey is complete.
- Perform ultra-stable measurements of the Sun and other celestial sources of radiation, avoiding the uncertainty introduced by radiation traversing the atmosphere. As an example, the ERB/ACRIM series and other instruments flown on a variety of satellites for the past 30 years have provided the current best measurement of total solar output.
- Perform measurements of the solar system and beyond using an observing baseline greater than the Earth's diameter. Spacecraft such as NASA's twin STEREO (Solar TERrestrial RELations Observatory) probes have been placed to observe the Sun from vantage points that are different from that available on Earth. Interferometric observations have not yet been undertaken with this sort of baseline.
- Analyze the space environment around the Earth (particles and fields)—more than 100 satellites and space probes have undertaken this type of measurement—to characterize the Earth-Sun system;
- Visit the planets, moons and other regions of the solar system, make in situ measurements (particles, fields, images, gravity, chemistry, geology, atmosphere, etc.) and/or return them to Earth. Spacecraft have now visited all the planets and several minor planets, asteroids and comets. All the planets except Uranus and Neptune have had spacecraft placed in orbit around them. Probes have landed on the surface of Mars, Venus and Saturn's moon Titan. Various samples have been returned to Earth from the Moon, while the Stardust space probe in 2006 became the first to collect and return samples of a comet, while passing through the coma of comet Wild 2, along with cosmic dust samples. Similarly, the Hayabusa space probe performed the first sample return from an asteroid in 2010, collecting and returning tiny grains of material from 25143 Itokawa. The Voyager-1 and Voyager-2 spacecraft reached the heliosheath region of the outer solar system in 2004 and 2007, respectively, where the solar wind slows from supersonic to subsonic speed, with Voyager-1 leaving the solar system on 25 August 2012;
- Measure cosmic rays directly—ground-based methods involve the detection of secondary or tertiary products of the collision of cosmic rays with the Earth's atmosphere. The 6¾ metric ton Alpha Magnetic Spectrometer (AMS-02) deployed on the International Space Station in 2011 is the most powerful instrument orbited for this purpose (at least since the Soviet's 16 ton Proton 4 that spent 8 months in orbit in 1968–1969), see [Sect. 1.2.1](#).

1.2.3.2 Earth Science

Space-based sensors can be an attractive way to undertake Earth science because of the following factors

- Viewed from space, large-scale geological, meteorological, hydrological, marine and biological phenomena become more apparent than when viewed from lower altitude.
- Hard-to-access regions of the world such as the polar regions can be extensively observed without compromising human safety.
- The top and upper regions of the atmosphere can be observed.
- The Earth's gravity field can be observed on a larger scale and more comprehensively than is possible on the ground.
- Observations over the oceans (for example of sea surface temperature) can be undertaken more comprehensively than is affordable using surface, airborne or sub-sea sensors.
- Altimetry from space provides information on sea level, glacier extent/volume, and similar phenomena that are impossible or very difficult to measure any other way.
- As new techniques are devised to measure phenomena of interest, space-based sensors provide a rapid means of achieving their deployment on a global-scale. Examples include atmospheric trace constituents, marine pollution and currents, and the thickness and age of ice.

Most space-faring nations have placed instruments in orbit that exploit some of the above factors. Climate change and other aspects of man's impact on the planet (deforestation, desertification, over-extraction of water, etc.) is driving many countries to increase the type and number of space-based sensors to monitor that impact [18].

1.2.4 National Security

Military space activities account for around one-fifth of the world's space economy. The activity is dominated by the US military, with typically more than double NASA's budget. The range of applications covers telecommunications, navigation, and various forms of surveillance and missile defense, each of which is discussed below. Monitoring of the threat from meteoroids and asteroids (Near Earth Objects, NEOs) is a special form of national security and is treated in this section.

1.2.4.1 Military Satellite Communications

The US, Russia, China, France, North Atlantic Treaty Organization (NATO), Germany, Spain and the UK deploy military communications satellites—the last three countries procure the services from a commercial operator. The military in these countries also procure extensive amounts of commercial communications satellite services—for the US and the UK, for example, the commercial satellite to military satellite ratio in terms of bandwidth used by the military forces is typically 4:1. The features of military communications satellites that differ from the commercial versions can include [19]

- Encryption
- Jam resistance (antenna nulling)
- Resistance to atomic explosions in space
- Rugged terminals
- Coding schemes to allow covert operation.

The armed forces of the developed world are increasingly operating in far-flung parts of the globe. The distance between the deployed forces in theater and the infrastructure at home makes satellites essential for modern military operations. Satellites are also essential for communications within theater as there will often not have been time to deploy a terrestrial communications network.

The nature of modern military operations demands more and more communications bandwidth. In recent years, the bandwidth requirement has risen fast because of the emergence of remotely controlled aircraft (UAVs or UASs). The smallest UAVs communicate to their operators only within the line of sight and thus do not require satellite links. However, the larger UAVs are the size of a small commercial airliner and satellite links are used to control them and to

obtain the surveillance data from them—for example, the US controls many of its larger UAVs in theater from Beale Air Force Base in California. The data sent back by these planes is often real-time video and thus very voluminous. Once the data is received in California, it is sent via another satellite link to the forces in theater that need it.

In such 'out of area' operations, satellites have to communicate to both large and small antennas. The field headquarters will typically have a relatively large antenna for communicating with home and will receive the large amounts of data that any modern organization needs for its operation. The individual patrols, whether in jeeps, tanks, ships, helicopters or aircraft, will each have a satellite link varying from a hand-held device for the soldier on foot to a steerable dish mounted on a vehicle. The mounting on a military vehicle typically has to be ultra-rugged to withstand the high accelerations (e.g. on a fast jet plane), rough treatment, dirt/sand/water/salt, heat (desert) or cold (polar/mountainous).

The combination of satellite communications and satellite navigation is central to avoiding 'Friendly Fire' (or, blue-on-blue) incidents. If every friendly unit is equipped with a GPS receiver to provide its exact location and a communications satellite terminal to send that information rapidly, covertly and reliably to field headquarters, a complete picture of friendly forces can be kept up-to-date. This function is sometimes called 'blue force tracking' or 'situational awareness'. Note that because of their high speed, aircraft have a separate dedicated system called 'Identification Friend or Foe' that uses line of sight (non-satellite) radio links.

1.2.4.2 Global Navigation Satellite System

The military in the US and Russia deployed the first two satellite navigation services—GPS (initially NavStar-GPS) and GLONASS (Russian: ГЛОНАСС), an acronym for Global Navigation Satellite System (Russian: Глобальная навигационная спутниковая система, or Globalnaya navigatsionnaya sputnikovaya sistema). Their objective is to provide real-time positioning information to personnel, equipment and weapons (missiles, bombs, artillery/mortar shells, etc.) worldwide. These systems are designed to enable the recipient of the GPS or GLONASS signals to establish their position without needing to transmit and thereby reveal their position to an adversary.

As discussed in Sect. 1.2.2 the GPS (and to a much lesser extent GLONASS) signals are the basis for a wide and growing community of civilian users. GPS transmits signals (the P-code) that can only be received by users who have specially equipped receivers—encryption techniques are used to prevent unauthorized users making use of the signals. In certain circumstances, over a specific region for example, the military authorities can turn off the unencrypted signals on the satellites and transmit only the military signals so that only their military will obtain positioning information.

Military receivers for GPS differ from civilian ones by virtue of the decryption facilities they contain. Receivers embedded in artillery shells, missiles, high-speed aircraft, etc. will also have special features to allow them to withstand the shocks and vibrations of their environment, and to counteract the Doppler effect of their high speed. GLONASS has similar features to GPS. Galileo will have a similar encrypted signal called the *Public Regulated Service* for which the decryption codes will be dispensed by European government authorities.

1.2.4.3 Military Surveillance Satellites: Weather Satellites

In some countries such as Italy the weather forecasting agencies are part of the military, whilst in countries such as the UK the weather forecasting agencies are located elsewhere in government but provides a service into the military.³ The weather satellite activities in those countries are therefore inherently both military and civilian. In other countries, such as the US, the military authorities have a weather forecast agency in addition to, and separate from, the civilian one. The US military weather agency has its own weather satellites called Defense Meteorological Satellite Program in low-Earth orbit. A plan to merge these satellites with the civilian NOAA series was canceled in 2009.

The features of military weather satellites that distinguish them from civilian ones are not published. It is known that at least some of their signals are encrypted to limit access to the data to authorized users.

1.2.4.4 Military Surveillance Satellites: Imaging Satellites

By 2013, eight countries acknowledged that they operated military imaging satellites: China, France, Germany, Israel, Italy, Japan, Russia and the US. In addition, the UK has one prototype satellite in orbit (TopSat). About 20 other countries operate surveillance satellites with sufficiently high resolution to be of interest to their military and which could therefore be classified as dual use (civil and military)—Italy's COSMO-SkyMed satellite constellation is explicitly dual use (30 % funded by the military).

Six of the eight countries mentioned above operate imaging radar satellites—France and Russia are the exceptions. France has an agreement with Germany and Italy to access their imaging radar satellites, while Russia has plans to deploy two imaging radar satellites, Kondor-E and Arcon-2.

Six of the eight countries operate optical imaging satellites—Germany and Italy are the exceptions. Thus China,

Israel, Japan and the US operate both optical and radar imaging systems.

The applications of these satellites are both tactical and strategic. At the tactical level, they provide information in support of military operations—providing maps for the preparation phase, detailing enemy dispositions during operations and providing feedback on results achieved after the event. Imaging radar is crucial in ensuring timely imagery independent of weather and time of day. Stereo optical imaging is offered by at least the US and French satellites, and is helpful in identifying objects. Multi-spectral imagery is provided by most of the optical systems, which is helpful in detecting subtle changes and countering camouflage.

Strategic applications include cataloging the assets and forces of a potential adversary including facilities for producing military equipment and weapons—uranium enrichment facilities are a well-publicized example of this type of target. Imaging satellites are typically in orbits about 400–800 km altitude, but some have a significantly lower perigee—there have been examples as low as 125 km in which case significant amounts of fuel have to be carried by the satellite to raise the perigee once an observing campaign has been completed. Many are in near-polar orbits to provide global coverage.

Military users want both wide area coverage and very high resolution. This incompatible combination is usually achieved by having two imaging systems (telescope/camera) on-board, although the US and Russia have sometimes placed these systems on separate satellites so that their missions can be undertaken independently.

The US and Russia have satellites in geostationary or another high altitude orbit to relay data from the imaging satellites back to base. Without such a relay facility, users might have to wait an hour or more until the imaging satellite passed over a friendly ground receiving station. China has launched a prototype version of a similar system [20].

1.2.4.5 Military Surveillance Satellites: Electronic Surveillance

The US, Russia and recently China operate fleets of satellites in orbits of about 1,000 km altitude that detect transmissions from below. These are especially valuable in monitoring the location of shipping since a ship in the middle of the ocean stands out strongly against the radio-quiet background. Some Russian satellites not only listen for signals but also use radar to seek ships and perhaps submarines. Many of these systems comprise two or three satellites (sometimes a 'mother' ship and one or more sub-satellites) separated by some tens of kilometers allowing the position of the objects below to be triangulated.

The US also operates a fleet of satellites in geostationary and other high altitude orbits that monitor radio communications and radar signals. Very little is known outside

³ Within the UK the weather forecasting agency, the Met Office, was an executive agency of the Ministry of Defence until July 2011 when it transferred into the Department for Business, Innovation and Skills.

authorized circles about these systems. A satellite believed to be of this type was launched in November 2010 and stated by the head of the US National Reconnaissance Office to be “*the largest satellite in the world*” [21]. This was taken by commentators to be a reference to a satellite that deploys a very large antenna once in orbit—thought to be five or more times the roughly 18 m diameter of the largest civilian antennas in orbit. The large antenna is thought to allow the satellites to receive from the Earth’s surface 36,000 km below the very weak signals emitted upwards by mobile phones and towers, terrestrial microwave towers, military radio telephones, telemetry from missiles and aircraft, et cetera [22].

1.2.4.6 Missile Defense

The US and Russia have satellites in geostationary and other high altitude orbits that detect the launch of missiles. France is experimenting with a satellite to detect missile launches and is known to be seeking collaborators among European Union member countries. The US Defense Support Program (DSP) satellites scan the Earth below every ten seconds and detect bright flashes. Details of the flashes are analyzed and those conforming to the pattern of a missile launch are reported—lightning flashes, forest fires, oil well explosions, meteors entering the atmosphere and other forms of illumination if detected are eliminated by analysis. The location of the flash is computed with an accuracy of about 15 km. By detecting a missile at various points in its trajectory, the location of its launch site can be estimated.

In May 2011, the first of a new series of US satellites called Space-Based Infrared System (SBIRS) was launched to geostationary orbit. However, two SBIRS sensors had already been launched by this point, hosted on two classified satellites in highly elliptical orbits. SBIRS is divided into SBIRS High, consisting of the two hosted sensors plus four geostationary spacecraft, and SBIRS Low, or Space Tracking and Surveillance System. SBIRS Low was originally expected to consist of about 24 satellites in low-Earth orbit. However, as of 2010 only two technology demonstration satellites had been launched. SBIRS is intended to replace DSP. The SBIRS satellites provide images of the scene below as well as duplicating the DSP functions. The DSP satellites provided information after the fact and were not intended to be part of a missile defense system, as is the case for SBIRS. The key additional feature needed for missile defense would be rapid dissemination of the detection of a missile launch so that countervailing actions could be initiated.

The US GPS satellites carry special sensors as a secondary payload to detect the explosion of a nuclear bomb designed to detect the characteristic double-peak flash of a nuclear explosion [23]. Prior to GPS these sensors were carried on dedicated satellites called Vela.

1.2.4.7 Near Earth Objects

Small objects from outer space hit the Earth all the time—shooting stars. Bigger objects do so less frequently. An object measuring about 40 m across struck Siberia in 1908 and exploded in the atmosphere causing damage roughly equivalent to that from the explosion of a ten-megaton nuclear bomb. Such objects are thought to hit Earth every few centuries. More recently, the airburst of the smaller, 17 m asteroid over Chelyabinsk, Russia in February 2013 was equivalent to a mere 0.44 megatons; about 1.8 PJ, or 20–30 times more energy than was released from the atomic bomb detonated at Hiroshima.

NASA has been tasked by the US Congress to identify 90 % of asteroids and comets in the inner solar system bigger than 140 m by 2020. A special camera is being installed on a mountaintop in Hawaii to scan the sky for these faint objects. The European Space Agency’s Gaia satellite will also help complete the survey. As a side effect of its main mission to map a billion stars in our galaxy, Gaia will identify and locate thousands of asteroids and comets. NASA’s Kepler planet finder satellite also provided a similar serendipitous service albeit without Gaia’s full-sky coverage [24].

1.3 Space as a Technology Incubator

New technology is continually being developed for and developed by the aerospace industry. These two avenues toward new technology development have many similarities, as well as many differences and are known as spin-in technology (also known as technology infusion) and spin-off technology. Spin-in technology can be thought of as bringing in a technology to solve a problem, while spin-off technology is the resulting technology developed to solve a problem which can subsequently be transferred to another arena, where the technology has use to solve new problems not envisioned when it was originally developed. Both avenues also involve innovation, and this innovation comes from a “*conscious, purposeful search for innovation opportunities*” [25].

The flowchart in Fig. 1.17 shows the process of technology infusion, which begins with an idea and a need. These parallel paths may or may not occur at the same time—often a technological innovation may not have an immediate need, or may rely on development of another technology. An idea generator is the technology creator. The user is the person with the technical problem. The idea generator must recognize that there are opportunities to use these ideas, while the user needs to translate their technical problems into requirements. While the ideal timing is to have both paths converge at the same time, this generally does not happen. Ideas are generated on their own schedule, and if problems are not identified, the idea may languish until an application is found, or it may lazily

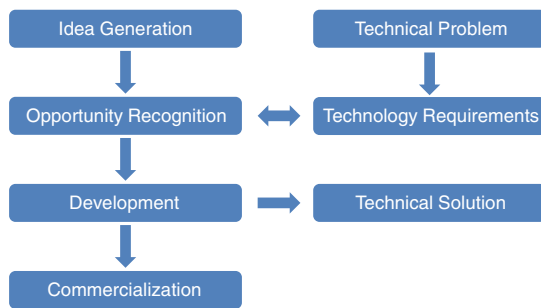


Fig. 1.17 Technology spin-in process

proceed to a finished concept. Avoiding the temptation for the user to reinvent the wheel can be difficult, and these users should exhaustively search for existing solutions to solve their problems. Once the connection is made between the idea generator and the user, the opportunity recognition and technology requirements phases are reached, and the two paths cross each other. This allows the space user to apply these ideas to solve their problems. The user can also identify a problem and can look for a solution from an idea generator. Should no ideas be found, the user can then choose to reassess their problems and move in another direction, or engage a technology developer to create a suitable technology.

Once the space user finds a promising solution, there may be significant interaction between the idea generator and the space user. Once the space user has determined that the idea generator can develop a solution to the technical problems, the idea generator (or someone else working in conjunction with the idea generator), with guidance from the space user, works with domain experts to apply their idea to the problem at hand. Upon completion of the idea development (which may be in the form of hardware, software or simply an analysis), the newly created technology is applied to solve a problem that was not envisioned when the idea was first created. The desired outcome of this process is to solve a technical problem; there may be a bonus outcome from the creation of a commercial product. The commercial product may or may not be for problems in the same industry—they may be for problems in a completely different field. Computers are an example technology where the concepts of microprocessors were developed independently from space applications, but later became an important technology infusion into space technology projects.

The second avenue is spin-off technology. The process is shown in the flowchart in Fig. 1.18. There are a significant number of similarities between the spin-off and spin-in process. Both are used as methods to support the solution of a problem. Where spin-off differs from spin-in is in the motivation for the process. As part of larger technology development programs, spin-offs generally occur following the development of a new technology that solves a problem.

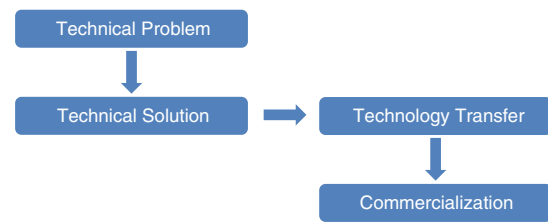


Fig. 1.18 Technology spin-off process

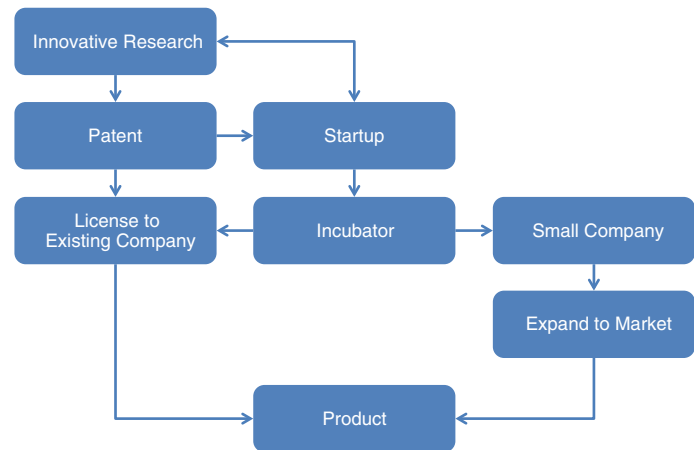
The process begins with the identification of problems in the course of developing a larger project. The project determines if there are available solutions that can be spun-into the project. Should none be found, a technology development process is initiated and the problem is (hopefully) solved. In the course of solving the technical problems, the outcome of the solutions could potentially result in a commercial product. An excellent example of a spin-off technology was the ARPANet computer network development for the US Department of Defense in the late 1960s. This network was originally developed for military computers to communicate with each other, and it ultimately grew into arguably one of the most successful commercial products from a spin-off technology in history—the Internet.

1.3.1 Technical Incubators and Commercialization of Space Technology

Many technology endeavors get their commercialization start as part of a technology incubator. The prime mission of technology incubators is to provide services to entrepreneurs who are interested in commercializing an idea [26]. They provide the entrepreneur with business support, marketing assistance, financial management, networking to strategic partners and investors, technology and intellectual property management. The flow of innovation to product is shown in Fig. 1.19. As universities tend to be a major source of technical innovation, many of these technical incubators are located at or near large research universities, and were created by the university. Other origins of technical incubators come from government agencies that are trying to attract high technology industry to the locale. These incubators have been working since the 1970s to lessen the burden of managing innovation from the entrepreneur and moving that into a management infrastructure that allows the entrepreneur to capitalize on the technology they are commercializing, and ultimately get it into a product. Fees for such incubators range from fee-for-service to financial stakes in the commercialized technology.

Virtually every high technology industry continues to benefit from technology incubators. Space technology

Fig. 1.19 Strategic map:
Innovation to product



development is no different, and has taken advantage of these incubators in various places throughout the world. A good example of how a technology incubator has fostered the spin-in of technology began during the 1970s, when several members of the Electrical Engineering Department at the University of Surrey in Surrey, UK, developed a paradigm to build satellites quickly and inexpensive. They started by building small satellites with commercial off-the-shelf components, and in 1985, the university formed Surrey Satellite Technology Limited (SSTL). By 2008 when the university sold its majority stake to EADS Astrium NV, the company had grown into a 300 person company and had developed (built or under construction) about 40 satellites. This move completed the process of spinning off a new space technology industry created through their technology incubation infrastructure.

1.3.2 Spin-In Technologies

Not all of the technology used in the space industry was developed specifically for the space industry. Many advanced technology developments that have been created for other applications have found their way into space programs. Some technology development activities for other fields, such as microelectronics and ballistic missiles, have been readily identified as having space applications, and structured pathways devised to infuse this technology into programs. Other technology development programs, such as photographic film, Teflon, or elastic fabric, are created without any thoughts of infusing this into the space industry, and some innovative group realizes how these specific technologies could solve a problem—they pair a solution that was developed for something else with a problem in the space industry.

Technology infusion can also be stimulated by an unmet need in a technology-based program. Many government-supported space and military agencies provide opportunities

for private industry and academia to either develop technology that can be infused into programs or support adapting technology for space applications. The US government, for example, has several programs in various agencies and departments that foster technology development. Two example organizations, NASA and the Defense Advanced Research Projects Agency (DARPA) each sponsor Small Business Innovative Research (SBIR). These programs provide seed money in a phase 1 program to develop new technology quickly. Promising technologies are further developed in a phase 2 program that provides funding over a longer time to mature the technology. Ultimately, the technology is infused into the sponsoring organization's programs along with a parallel path towards commercialization. Many companies have similar programs to foster the development of new technologies, and generally provide funding as part of a venture capital or equity sharing arrangement. Unlike governments, whose primary motivation is to support technology development to fulfill their mission, companies and investors are motivated by rapid commercialization and favorable return on their investment.

1.3.3 Spin-Off Technologies

Technological innovation for the space industry has become the source for many every-day products used throughout the world. While the motivation for the innovation behind these items has been to solve immediate problems in the space industry, the opportunities to spin these technologies into commercial products range from being obvious, to some bright innovator seeing a not-so-obvious application.

1.3.3.1 Sources of Spin-Off Space Technology

Spin-off of space technology is woven into so many different sectors of daily lives, in both obvious and not-so-obvious applications. Whilst it is apparent that space technology has come from many sources over the years, it is difficult to put a

Fig. 1.20 Landing a spacecraft on Earth or a space probe on Saturn's moon Titan (as seen on the left), is just like dropping a potato crisp into a bag. Image ESA (left) and Amos (right)



monetary value on how much society has benefitted from space technology.

Governments and industry throughout the world have recognized the benefits of spin-off technology. In addition to the US government's technology transfer activities in organizations such as NASA and DARPA in the US, other countries are engaging in the spin-off of their space technology. Examples include the National Research Council Canada's Industrial Research Assistance Program which is charged with supporting Canadian industry. The Russian Technology Transfer Center coordinates technology transfer between the Russian aerospace and other high technology companies and both the Russian and international markets. The European Space Agency's Technology Transfer Program is tasked with highlighting benefits of the European space program and strengthening the competitiveness of European industry by infusing ESA-developed technology into member's industry. Many of the other space-faring nations have similar activities to infuse home-grown technologies into both domestic industry and the international market.

1.3.3.2 Examples of Spin-Off Space Technology

Spin-off technology from space programs are imbedded throughout modern society. These technologies have improved standards of living throughout the world, and continue to do so. Various breakthroughs have improved life expectancies, made instantaneous communication anywhere in the world possible, and have made cost-effective and efficient travel available to many people. Most people don't realize how widespread the infusion of space technology into their lives is, but nearly every area of advanced technology has some connection with space technology.

While there are innumerable examples of spin-off space technology, an interesting example of a spin-off that is about as far from the space technology field as it is possible to imagine is in the snack food industry. In 2009, the European Space Agency's Technology Transfer Programme awarded

their first 'Space Spin-Off' award for outstanding performance in the area of space to non-space technology transfer to HTG Hyperschall Technologie Göttingen (Germany) for applying technology developed for soft landing a spacecraft to packaging potato crisps (chips).⁴ HTG realized that this soft landing technology, developed by ESA, could be applied to machines that package delicate potato crisps. Applying this technology, their new filling machine resulted in increases of 30 to 50 % in speed (Fig. 1.20).

1.3.4 Future Opportunities

Future opportunities to infuse technology into the space industry and spin technologies out of the space industry are, of course, difficult to predict. Breakthroughs can happen with little notice. Unexpected outcomes can spawn new applications, new products and even new industries. Predicting the future is difficult, at best, and any attempt to predict future development from space technology will be fraught with errors. Predictions from the 1950s, such as flying cars in every garage, have shown us the limitations of forecasting the future. On the other hand, predictions of personal communicators, seen in science fiction of the 1960s such as *Star Trek*, have materialized with the widespread availability of cell phones.

Rather than trying to predict the future, the more accurate prediction would be to make sure that the environment is right for spawning future technologies. The environment to exploit spin-in and spin-off technologies must be favorable for innovators. Inventors will be more likely to develop innovative technology if the pathway to commercialization is fairly simple through minimization of the bureaucratic burden and strong legal protection of the intellectual property.

⁴ ESA Technology Transfer Programme, "Company wins spin-off prize for keeping potato crisps intact", 27 April 2009, <http://www.esa.int/ttp>, date cited May 2011.

References

1. Tsiolkovsky, K., 'Investigation of outer space rocket appliances (Исследование мировых пространств реактивными приборами)', The Science Review, Volume 5, 1903.
2. Moore, W., "A Treatise on the Motion of Rockets: To which is added an Essay on Navel Gunnery, in Theory and Practice; designed for the use of the Army and Navy, and in all places of Military, Naval and Scientific Instruction.", Printed for G. and S. Robinson, Paternoster-row, London, 1813.
3. Coquilhat, C. E., "Trajectoires des fusées volantes dans le vide", Mémoires de la Société Royale des Sciences de Liège, Vol. 5, November 1873. In French.
4. Report from the Commission to the European Parliament and the Council, 'Mid-Term review of the European satellite radio navigation programme' COM(2011) 5 Final, Brussels, January 2011.
5. Webb, J.E., McNamara, R.S., 'Recommendations for Our National Space Program: Changes, Policies, Goals', May 1961. Contained within "Exploring the Unknown", NASA SP-4407, 1995.
6. "National Space Policy of the United States of America", White House, Washington DC, 28 June 2010, p10.
7. "Reference Guide to the International Space Station", NASA, Washington DC, NP-2010-09-682-HQ, Assembly complete edition, November 2010.
8. Norris, P., "Weather Satellites", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 23-44.
9. Kelly, G., Thépaut, J.-N., "The relative contributions of the various space observing systems", *Joint 2007 EUMETSAT Meteorological Satellite Conference and the 15th Satellite Meteorology & Oceanography Conference of the American Meteorological Society*, Eumetsat, Darmstadt Germany, P.50, 2007.
10. "Cospas-Sarsat Report on System Status and Operations No. 28, January – December 2011", C/S R.007, Annex C, October 2012.
11. "The Space Report 2013", The Space Foundation, Washington DC & Colorado Springs CO US, 2013.
12. Taverna, M. A., "End of the Cycle", *Aviation Week & Space Technology*, 15 March 2010, p 48-50.
13. Taverna, M. A., "Bullish on Broadband", *Aviation Week & Space Technology*, 15 March 2010, p.51-52.
14. Taverna, M. A., "New Wrinkle", *Aviation Week & Space Technology*, 7 March 2011, p.90-93.
15. Lardier C., "Satellites de télécoms: l'arrivée de la large bande", *Air & Cosmos*, 11 March 2011, p 20-25.
16. Norris, P., "Where am I, where are they?", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 131-143.
17. Norris, P., *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp56-57, 73-85, 124-125, 197-218.
18. Norris, P., "Climate Change, *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 45-69.
19. Butler, A., "Fast and Secure", *Aviation Week & Space Technology*, 7 April 2008, pp 52-54.
20. Norris, P., "Military Imaging Satellites", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 189-232.
21. Carlson, B., "National Reconnaissance Office Update", Air Force Association Air & Space Conference, National Harbor MD, 13 September 2010.
22. Norris, P., "Military Radio Surveillance from Space", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 233-260.
23. Norris, P., "Missile Early Warning", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 218-228.
24. Norris, P., "The Asteroids", *Watching Earth from Space*, Springer Praxis, Chichester UK, 2010, pp 69-72.
25. Drucker, P. F. "The Discipline of Innovation", Harvard Business Review, pp5-10, August 2002.
26. Knopp, L. "State of the Business Incubation Industry, 2006", Athens, Ohio: National Business Incubation Association, 2007.

Further Reading

27. Bekey, I., "Advanced Space System Concepts and Technologies", AIAA, 2003. ISBN: 978-1884989124.
28. McCurdy, H. E., "Space and the American Imagination", The Johns Hopkins University Press, 2nd edition, 2011. ISBN: 978-0801898679.
29. Norris, P., "Spies in the Sky: Surveillance Satellites in War and Peace", Praxis, 2007. ISBN: 978-0387716725.
30. Norris, P., "Watching Earth from Space: How Surveillance Helps Us – and Harms Us", Praxis 2010. ISBN: 978-1441969378.
31. "Reference Guide to the International Space Station", NASA, Washington DC, NP-2010-09-682-HQ, Assembly complete edition, November 2010.
32. For communications satellites, the annual "International Communications Satellite Systems Conference (ICSSC)" provides up-to-the-minute information (mainly non-military).
33. For navigation satellites, the annual Institute of Navigation (ION) GNSS conference provides the most authoritative English language non-military information.
34. For space science, the annual COSPAR conference provides a global update.
35. For space exploration, the annual International Astronautical Congresses provides wide-ranging and global updates.

Malcolm Macdonald

Until well after World War II, with no meaningful definition distinguishing the terminology, the fields of aerospace, aeronautics and astronautics were synonymous. Indeed even today, the terms are widely misunderstood and misused. However, the great Hungarian aeronautical engineer and physicist, Theodore von Kármán (original Hungarian name: Szöllőskislaki Kármán Tódor; 1881–1963) believed a clear distinction between aeronautics and astronautics could, and should be made. Therefore, in the early 1950s, and in consultation with the International Federation of Astronautics (IAF), founded 1951, and the Fédération Aéronautique Internationale (FAI), von Kármán undertook the task of defining the respective terms.

In aeronautics, the presence of an atmosphere is critical, while in astronautics its absence is critical. As altitude is increased the atmospheric density decreases. Thus, for steady level flight, controlled by aerodynamic forces, the velocity of the vehicle must increase until eventually the required velocity will overcome the circular orbit velocity. Hence, aerodynamic forces are no longer required to maintain steady level flight. The converse is true for astronautics. As altitude is decreased, the notion of a free-fall orbit becomes meaningless due to the increasing atmospheric density, leading to an increase in the drag force. In conclusion, von Kármán and his co-workers determined that the nominal boundary could be set at an altitude of around 100 km, a definition readily accepted by the IAF. Meanwhile the FAI, who to this day administrate aeronautics records and hence had a slightly different interest in the definition, created a new category of flying machine, named spacecraft, which from that point on would have separate records to aircraft. Section 8 of the FAI Sporting Code

would, thereafter govern such machines, and the distinction between aeronautics and astronautics.¹ The code defines the nominal boundary to space as the von Kármán ellipsoid, an ellipsoid at 100 km altitude; often termed simply as the Kármán line. A spacecraft is thus a vehicle or vessel designed to operate beyond the von Kármán ellipsoid. By extension of this definition, crafts such as rovers, landers or (non-Earth) atmospheric probes are also termed spacecraft. Note that the plural of spacecraft is spacecraft.

Having established a simple and clear definition of a spacecraft, reality must unfortunately intervene. Within the space community, the term spacecraft has two contradictory meanings in common parlance. The first refers to the spacecraft as the whole vehicle, while the other refers only to the platform onto which the payload is mounted. For this reason, the term satellite is often used, a term which simply means a body orbiting another of larger size. However, not all spacecraft orbit and hence the terms space probe or space vehicle can be used when satellite is inappropriate, such as a Mars lander. Within this book, all of these terms are used in-line with in common parlance (Fig. 2.1).

Just as the von Kármán ellipsoid is not actually a hard and clear boundary between aircraft and spacecraft, space technology cannot be considered solely as the space vehicle, rather the vehicle is part of a much larger system. A space system can be considered the entirety of hardware, software and human resources required to conduct a space mission. The space system is typically subdivided into the space segment and the ground segment.

The space segment is the spacecraft, while the ground segment is the system on Earth that manages and controls the spacecraft, and its data products. The ground segment can be subdivided into two core components; the flight operations segment, relating to the spacecraft housekeeping,

M. Macdonald (✉)
Advanced Space Concepts Laboratory, Strathclyde Space
Institute, University of Strathclyde, Glasgow, Scotland
e-mail: malcolm.macdonald.102@strath.ac.uk

¹ See www.fai.org.



Fig. 2.1 Image of the upper regions of the Earth's atmosphere, from approximately 28 km altitude, leading to space and including the region of the von Kármán ellipsoid. *Image* University of Strathclyde

or telemetry data, and commanding, and the payload data ground segment, relating to the spacecraft data product. The flight operations segment will typically be managed by a single control center. However, this center may itself be supported by other secondary centers. The spacecraft control center is ultimately responsible for the safe operations of the spacecraft. Moreover, under nominal operations it will be the sole originator of all spacecraft commands.

The spacecraft's data product can be disseminated in many ways, typically defined by the spacecraft mission, as shown in Fig. 2.2. It should also be noted that the overall architecture need not include a direct-link from spacecraft to ground, but can use an inter-spacecraft link, as also shown in Fig. 2.2.

The final component of the space system is the launch vehicle, which has the primary objective of traversing the von Kármán ellipsoid to deliver a payload, i.e. a space vehicle, into space. The launch vehicle need not specifically establish its payload in an Earth orbit, rather it can enter a suborbital, or parabolic arc, it can place the payload directly onto an Earth escape trajectory, perhaps *en route* to another planet, or it can place it into an Earth orbit. The final orbit of a spacecraft is often actually achieved through a combination of the launch vehicle and the spacecraft's own propulsive capabilities. For example, a geostationary communications spacecraft is typically inserted by the launch vehicle into a geostationary transfer orbit (GTO) with apogee at geostationary distance, see Chap. 4, and perigee at only a few hundred kilometers altitude. The communications spacecraft will thereafter use its own propulsive capabilities to maneuver into a geostationary orbit (GEO). Thus, the functional boundary between the final stage of the multi-stage launch vehicle and the propulsive capabilities of the launch vehicles payload is somewhat ambiguous. As such, within this handbook space

transportation systems are considered simply as a different type of spacecraft mission objective or phase.

2.1 The Space Segment

The space segment is defined as everything beyond the von Kármán ellipsoid. As shown in Fig. 2.2 the space segment architecture can take different forms, perhaps with spacecraft providing services to other spacecraft in a manner which may, or may not, have been envisaged when either spacecraft was commissioned. Most typically, such services include communications or navigation assistance.

The space segment can also be constructed of several spacecraft working in isolation, and largely operated as individuals, to provide a coherent ground-segment data product; this is termed a spacecraft constellation. Several spacecraft constellations are in service today. Perhaps the most widely known of these is the global navigation satellite system (GNSS) maintained by the United States government, under the stewardship of the Department of Defense, as a national resource, called the Global Positioning System (GPS). Historically, the other principal GNSS system was the Russian GLObal Navigation Satellite System (GLONASS), which was used solely by the Russian military until 2007, when it was made available to civilians. However, as discussed in Chap. 1 several other nations are now keenly pursuing this technology, including the Chinese BeiDou-2 navigation system and the European Union's Galileo positioning system. It is of note that many spacecraft today use GPS to aid in-orbit navigation. Another spacecraft constellation of note is the Iridium constellation, owned and operated by Iridium Communications Inc., consisting of over 60 spacecraft providing voice and data coverage to satellite phones, pagers and integrated transceivers over Earth's entire surface. A key feature of the Iridium constellation, and all other space-backbone mobile phone systems, is the ability to operate in areas of limited infrastructure, making them of significant value not only to the military, but also in disaster relief efforts where the infrastructure has been destroyed (Fig. 2.3).

Alternatively, spacecraft can work co-operatively to form a single integrated space segment, this is termed formation flying and quite a few natural formations are possible, see Chap. 4. Indeed, several spacecraft are claimed to have flown in formation, for example, ESA have previously flown the ERS-2, European Remote-Sensing Satellite-2, spacecraft and ENVISAT, Environmental Satellite, in a tandem formation enabling synthetic aperture radar (SAR) interferometry, or InSAR measurements to be made. InSAR combines two or more SAR images of the same site to allow slight variations that may have occurred between image acquisitions to be detected. As shown in Fig. 2.4, the

Fig. 2.2 The generic space system (not to scale); comprising the space segment and the ground segment. *Image Malcolm Macdonald*

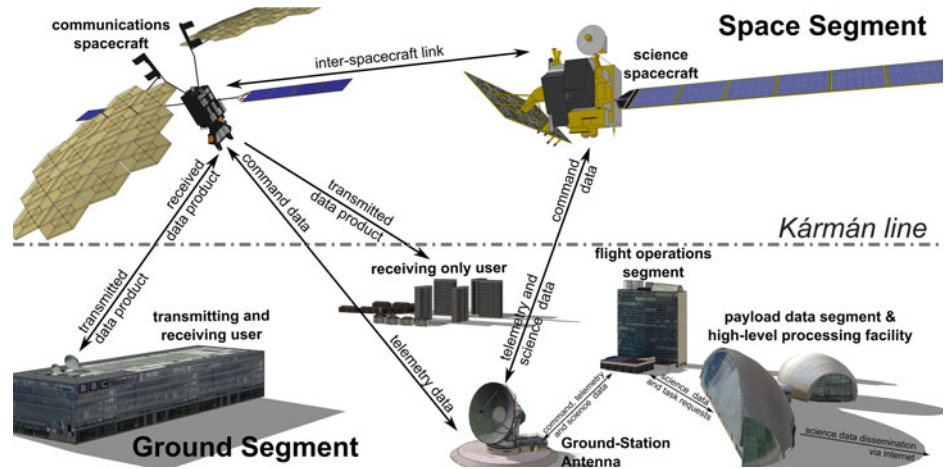
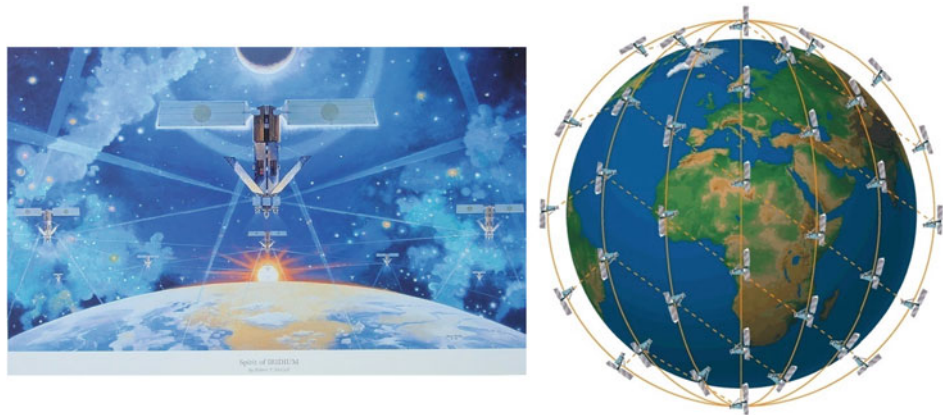


Fig. 2.3 An early Iridium poster (left) and the Iridium constellation (right). *Image Iridium Communications Inc*



ERS-ENVISAT tandem formation was configured such that SAR images of the same site would be acquired 28 min apart, enabling rapid variations to be detected. Figure 2.4 shows a sea ice displacement map acquired by the ERS-ENVISAT tandem formation where sea ice displacements of over 150 m were detected in less than half an hour. It should be noted however that the ERS-ENVISAT tandem formation is really closer to a two spacecraft constellation than a formation. Formation flying is perhaps best illustrated by mission concepts where the spacecraft are required to act in a coordinated manner in order to provide the required data product. Examples of this are the joint ESA/NASA Laser Interferometer Space Antenna (LISA), mission concept, or ESA's free-flying X-ray observatory mission concept, Xeus, where the mirror and detectors would be located on separate spacecraft, flying in formation 50 m apart.

2.1.1 Payload

For space science missions the payload is typically a bespoke suite of instruments. Meanwhile for commercial spacecraft, such as communications platforms the payload,

and its supporting platform, will typically have some significant flight heritage and may be produced many tens of times. However, it is easily forgotten by the spacecraft engineer that the payload is the *raison d'être* of any spacecraft. Indeed, the Merriam Webster Dictionary gives a particularly adept definition of payload as “*the load carried by a vehicle exclusive of what is necessary for its operation; especially: the load carried by an aircraft or spacecraft consisting of things (as passengers or instruments) necessary to the purpose of the flight.*” In other words, the payload is the biological passengers, or the part of a robotic vehicle that produces revenue, a product or a service. The principal purpose of the rest of the spacecraft is thus to serve the needs of the payload, positioning it where it needs to be in space, while providing it with power, communications and the desired thermal environment, whilst also ensuring it is pointing in the correct direction on a sufficiently stable platform.

It should be noted that the term payload is often used at various levels of the space system to denote different things; typically, this can be understood by considering the purpose of the vehicle. For example, the launch vehicle payload is the spacecraft, while the spacecraft may have a payload that

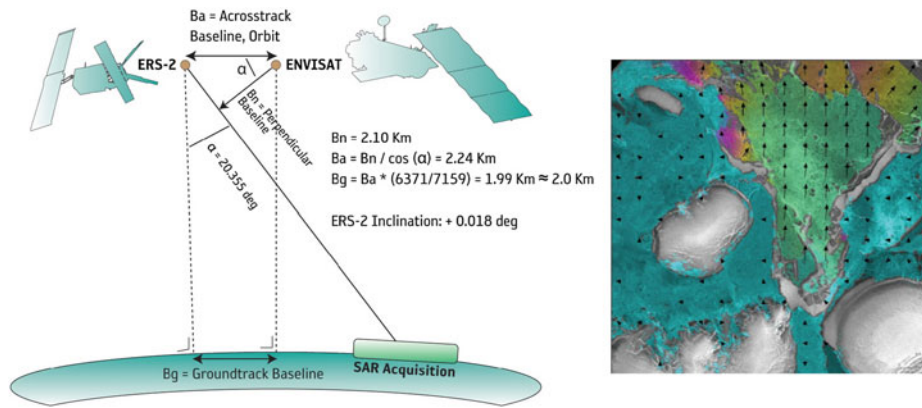
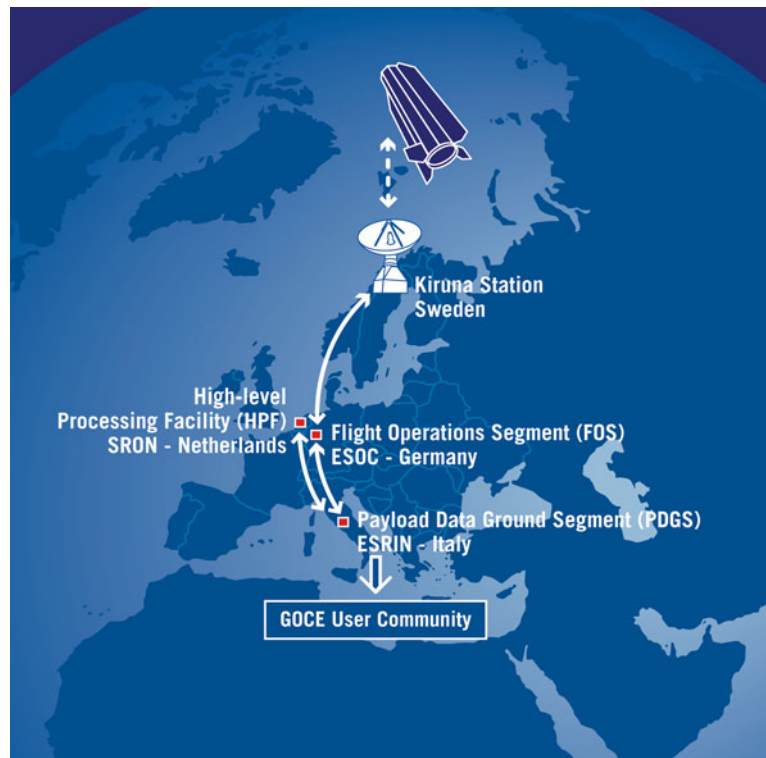


Fig. 2.4 Geometry of ERS-envisat tandem operation (*left*) and geocoded sea ice displacement map (*right*); the *green* areas correspond to an observed sea ice displacement of about 160 m in 28 min. The

image brightness corresponds to the backscattering of the Envisat image. *Image* ESA

Fig. 2.5 The gravity field and steady-state ocean circulation explorer (GOCE), mission space system and data flow. *Image* ESA—AOES Medialab



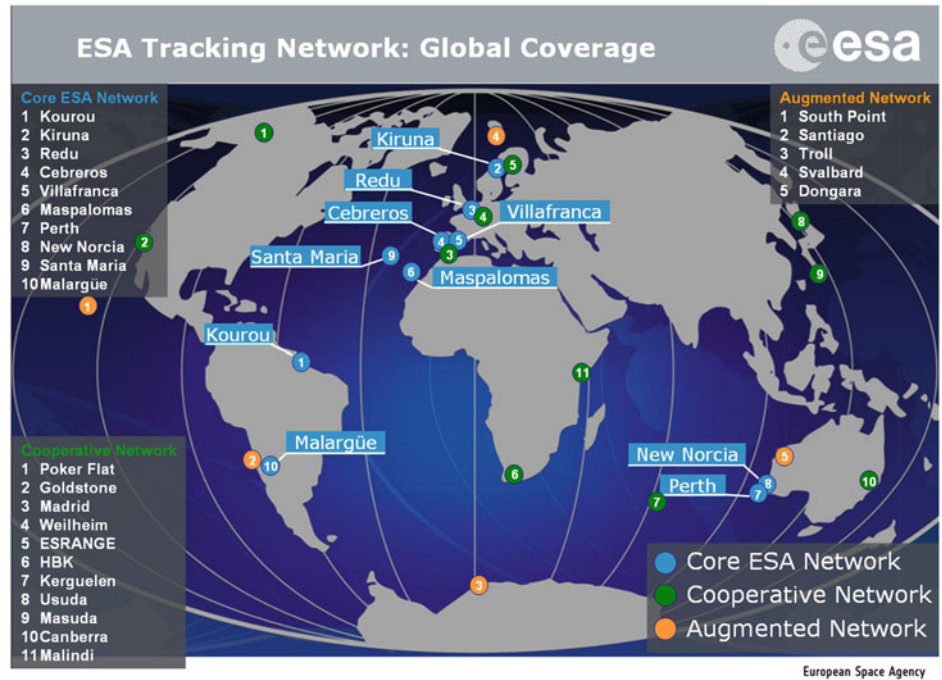
is a communications system, science instruments, or, say, a lander. The lander then may have a science suite on board, but it may also carry a payload of a rover, and the rover may in turn have a science suite payload.

2.2 The Ground Segment

The ground segment is defined as everything before the von Kármán ellipsoid and consists of the entirety of hardware, software and human resources required to manage and

control a space vehicle. As discussed above, the ground segment can be subdivided into two core components, the flight operations segment and the payload data ground segment. The flight operations segment is relatively independent of the spacecraft mission, and is focused on the command and control of the spacecraft. However, the payload data ground segment is heavily defined by the mission objectives and the data product. For example, in a science mission the primary spacecraft control center will typically receive the flight operations data as well as the science data product. The data product will then be passed

Fig. 2.6 The ESA tracking network in January 2011. *Image* ESA



to the payload data ground segment, which may or may not be collocated. The payload data ground segment will then pass the data product to a science principal investigator (PI) for some initial high-level processing prior to the data being distributed widely, typically via the Internet as shown in Fig. 2.2. Furthermore, in a science mission the request for specific data products will also be managed by the spacecraft control center, as shown in Fig. 2.2. The ESA Gravity field and steady-state Ocean Circulation Explorer (GOCE), mission space system and ground-segment data flow is shown in Fig. 2.5.

Alternatively consider, for example, a communications spacecraft, where the flight operations segment will typically not be directly concerned with the data product, as shown in Fig. 2.2, and may in fact be wholly separate. Indeed, typically commercial data products, such as Direct-to-Home television, or mobile phone communications, are depended on this type of space system and ground segment architecture.

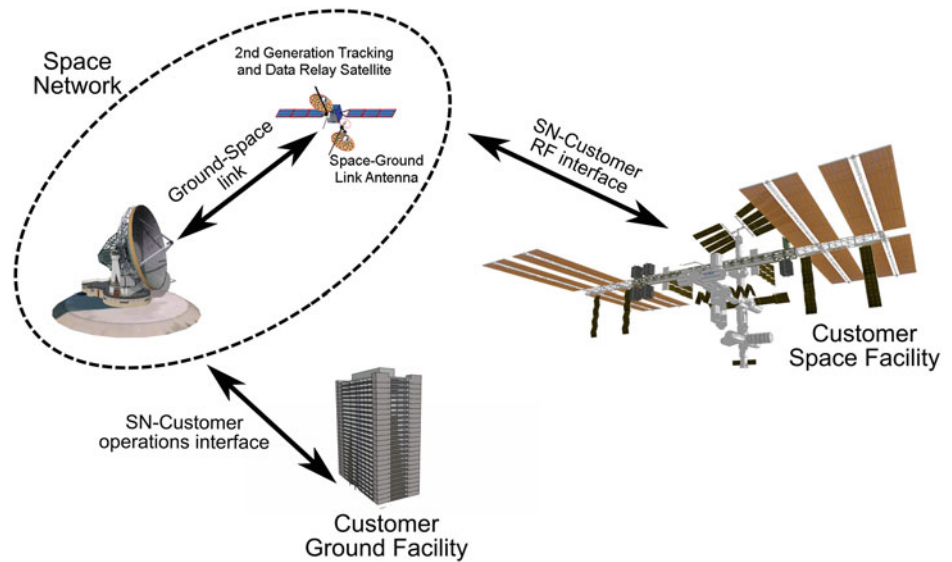
2.2.1 Ground Stations

To provide high quality, reliable and robust communications with spacecraft it is typical to use multiple ground stations, often positioned at geographically strategic locations. For low inclination spacecraft, the ground stations will ideally be distributed in longitude to ensure at least one communications window per revolution. While for polar orbiting spacecraft, ground stations close to the poles will provide one communications window per revolution.

Two well-known examples of ground station networks are ESA's tracking station network (ESTRACK), a worldwide system of ground stations providing links between spacecraft and ESA's Operations Control Centre at ESOC, and, the NASA-JPL operated Deep Space Network (DSN), which supports both Earth orbiting spacecraft and interplanetary missions. Note that the DSN is separate from NASA's Near Earth Network (NEN), which provides orbital communications support for Earth orbiting platforms via various NASA ground stations and is operated out of the Goddard Space Flight Center. The ESTRACK network is shown in Fig. 2.6.

It is perhaps a sign of the maturity of robotic space technology that the traditional divide between ground and space segment is, perhaps most notably disappearing when discussing ground station system architectures. An example of this is NASA's Space Network (SN) project established in the early 1980s to replace NASA's worldwide network of ground tracking stations. SN provides communications support to Earth orbiting spacecraft, such as the International Space Station, using both a traditional ground segment and a space segment, through geostationary Tracking and Data Relay Satellites (TDRS). SN can provide tracking and data acquisition services over 100 % of a spacecraft's orbit for altitudes between 73 and 3,000 km. The SN architecture is shown in Fig. 2.7, where it is seen that the traditional ground-segment is, in effect, being extended into the space segment. Note the proposed European Data Relay Satellite (EDRS) system, also mentioned in Chap. 1, is a further example of this type of extension.

Fig. 2.7 Space network customer and operations interface. *Image* Malcolm Macdonald



2.2.2 Operations

The operation of a spacecraft is often the only part of the space system which directly involves humans, other than of course human space flight. The operations team is the fundamental human element, integrating the system and the mission. Success will often depend on the quality of this team. As such, the operations team will develop carefully considered and detailed operations procedures, documents and manuals, and will train ahead of launch using an operations simulator. The operations simulator will also be used in-flight to check spacecraft commands prior to actually sending them to the spacecraft. The operations team of the Mercury Sigma-7 spacecraft is seen in Fig. 2.8, training in the control room prior to launch. Meanwhile, the operations team of CryoSat-2 is similarly seen in training almost 50 years later in the same figure. It should also be noted that the operations team extends significantly beyond the control room, to include support and specialist engineers, scientists and technologists, hardware and software support as well as general project, site and administrative support.

2.2.3 Two-Line Elements

A key objective of the ground segment is to determine the orbital ephemeris of the spacecraft. The Keplerian orbital parameters, see Chap. 4, can be encoded in a number of formats, but the most commonly used is the NORAD (North American Aerospace Defense Command) ‘Two-Line Element’, TLE, format due to its concise nature. The orbital ephemeris of many thousands of space objects, including both active spacecraft and orbital debris, is determined by NORAD, and freely distributed via the Internet in the form

of TLEs.² Two-Line Elements can easily be automatically retrieved for use in spacecraft trajectory simulation software. A sample TLE is shown in Table 2.1, where it is seen that the TLE consists of a title, followed by two lines of formatted text. From Table 2.1 it is seen that the International Space Station is in an orbit inclined 51.6° to the equator, completing 15.7 revolutions per day in a virtually circular path. Note that the BSTAR term in column 54 of line one of the TLE is an adjusted value of the ballistic coefficient, see Chap. 4, where the ballistic coefficient is multiplied by half of a reference value of atmospheric density.

2.3 Space Project Planning, Implementation and Technology

The space project begins with a set of top-level objectives, for example, the GOCE mission, launched in March 2009, had the objective to measure the Earth’s gravity field, and model the geoid with an unprecedented accuracy and spatial resolution. The mission analysis and design process then defines the space system, considering system and technology constraints, to define measurable mission objectives and metrics that can be achieved within the ultimate mission constraint of cost.

Several tools, methodologies and standards are available to the space system engineer to facilitate the process of mission analysis, design and technology assessment. Some of these are introduced here.

² See <http://celestrak.com/>.

Fig. 2.8 View of mercury control center, September 10, 1962, prior to the Mercury-Atlas-8 (MA-8) flight of the Sigma-7 (*top*; Photo IDs: S62-05139 and KSC-62PC-128), and the CryoSat-2 Mission Control Team in Main Control Room ESA-ESOC, December 8, 2009 (*bottom*; ID Number: SEMTLKOJH4G). Image NASA and ESA



2.3.1 ECSS: European Cooperation for Space Standardization

The European Cooperation for Space Standardization (ECSS),³ was established in 1993 to develop a coherent and definitive set of standards for use in all European space activities. Despite being intended as a European initiative, ECSS has gained a global importance and provides an excellent resource for the development of good practice. The ECSS standards are typically mandated for use in ESA missions and users are encouraged to provide feedback on usage to ensure the standards remain ‘live’ documents.

The ECSS documentation architecture contains three branches, these are ‘Management’, ‘Product Assurance’ and

‘Engineering’, each of which contains a subset of standard documents split into four hierarchical levels, defined to the detail level of detail required to differentiate major functions, disciplines and activities. These four levels are defined as

- *Level 0 (ECSS-P-00)*—describes the policy and objectives of the ECSS system and its architecture together with the principal rules for the creation, validation and maintenance of documents.
- *Level 1 (ECSS-M-00, ECSS-Q-00, ECSS-E-00)*—describes the strategy in the specific domain, gives a global view of the requirements, and outlines the interfaces between the elements (and the documents) at Level 2.
- *Level 2 (ECSS-M-10, ECSS-Q-10 ...)*—describes the required objectives and functions for all aspects in the individual domain (project organization, quality assurance, system engineering, etc.).

³ See www.ecss.nl.

Table 2.1 The TLE of the International Space Station on April 4 (day 94), 2011

International Space Station Two-Line Element			
ISS (ZARYA)			
1	25544U	98067A	11094.38711506 .00060886 00000-0 44580-3 0 1260
2	25544	51.6466	179.6373 0002360 82.3471 6.0048 15.72587753709286
Column	Characters	Description	Example
<i>Title Line</i>			
1	24	Satellite Name	ISS (ZARYA)
<i>LINE 1</i>			
1	1	Line No. Identification	1
3	5	Catalog No.	25544
8	1	Security Classification	U
10	2	International Identification (last two digits of launch year)	98
12	3	International Identification (launch number of year)	067
15	3	International Identification (piece of launch)	A
19	2	Epoch year (last two digits of)	11
21	12	Epoch day (day of year and fraction of day)	094.38711506
34	10	First time derivative of mean motion, divided by two	.00060886
45	8	Second time derivative of mean motion divided by six, decimal point assumed	00000-0
54	8	BSTAR drag term, decimal point assumed	44580-3
63	1	'Ephemeris type', now just the number 0	0
65	4	Element number	126
69	1	Checksum (modulo 10)	0
<i>LINE 2</i>			
1	1	Line No. Identification	2
3	5	Catalog No.	25544
9	8	Inclination	51.6466
18	8	Right Ascension of Ascending Node	179.6373
27	7	Eccentricity with assumed leading decimal	0002360
35	8	Argument of the Perigee	82.3471
44	8	Mean Anomaly	6.0048
53	11	Revolutions per Day (Mean Motion)	15.72587753
64	5	Revolution Number at Epoch	70928
69	1	Check Sum Modulo 10	6

- *Level 3*—describes methods, procedures and recommended tools to achieve the requirements of Level 2 documents. In addition, it defines the constraints and requirements for interfaces, and the performance of the specified product or activity. The Level 3 documents are guidelines and are allowed to be adapted to the needs of a project.

2.3.2 Project Phasing

The ECSS divides the space mission project life cycle into seven phases; these are defined in Table 2.2 alongside the equivalent six NASA phase definitions. It should be noted

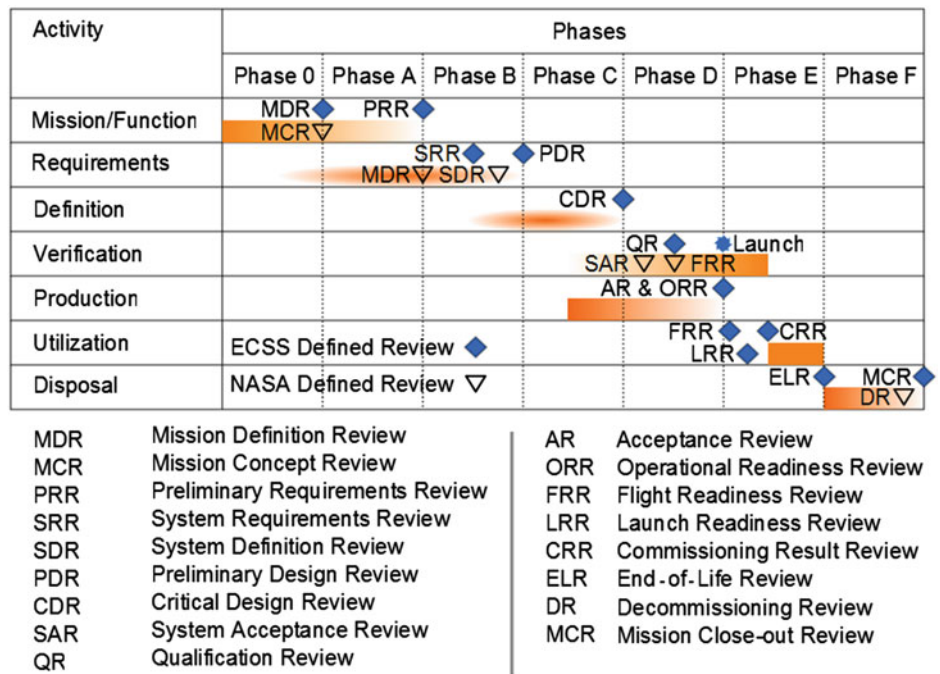
that other established space institutions, such as the US Department of Defense, often use their own project life cycle phasing.

Each project phase is associated with certain activities and project milestones, typically in the form of project reviews, which will also likely be payment milestones. The basic activities during each mission phase are illustrated in Fig. 2.9, where the ECSS-defined milestones are given alongside additional NASA-defined milestones. Note from Fig. 2.9 that on occasion the same review will be given a different name by ECSS and NASA. A detailed description of each mission phase can be found in the ECSS documentation;

Table 2.2 Space mission project life cycle phases as defined by ECSS and NASA

Phase ID	Phase name		
ECSS	NASA	ECSS	NASA
0	Pre-A	Mission analysis/needs analysis	Advanced studies
A	A	Feasibility	Preliminary analysis
B	B	Preliminary design	Definition
C	C	Detailed design	Design
D	D	Qualification and production	Development
E	E	Utilization	Operations
F		Disposal	

Fig. 2.9 A typical space mission life cycle with ECSS and NASA defined milestones. *Image* Malcolm Macdonald



see ECSS-M-ST-10C Rev. 1, “Project planning and implementation”, and will be discussed in more detail in Chap. 7.

2.3.3 TRL: Technology Readiness Level

The concept of ‘Technology readiness level’ (TRL), is used widely in aerospace to assess and define the maturity of a technical concept, capability or product. Nine technology readiness levels are defined and shown in Fig. 2.10, along with a more detailed, but NASA-centric, tabular definition in Table 2.1.1.

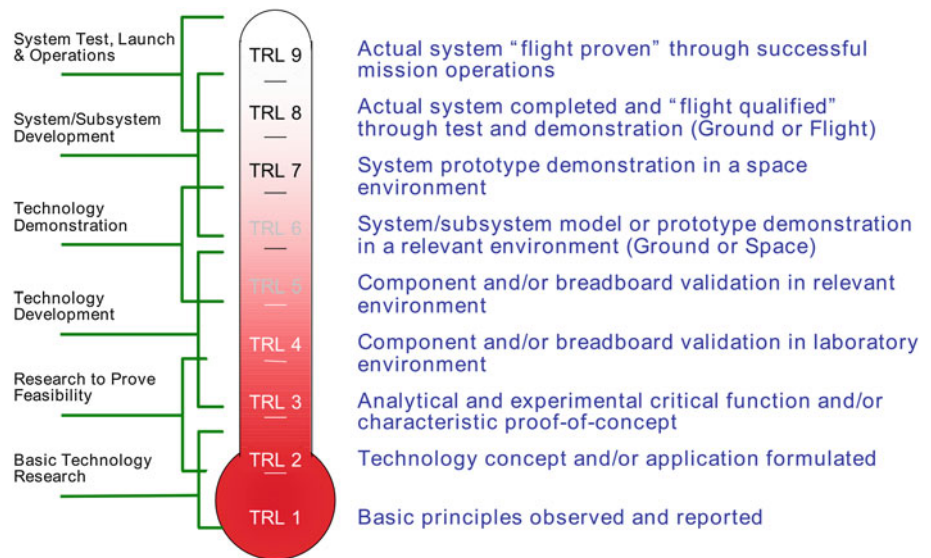
The technology readiness levels can be defined further as TRL 1. *Basic principles observed and reported*: Transition from scientific research to applied research. Essential characteristics and behaviors of systems and architectures. Descriptive tools are mathematical formulations or algorithms.

TRL 2. *Technology concept and/or application formulated*: Applied research. Theory and scientific principles are focused on a specific application area to define the concept. Characteristics of the application are described. Analytical tools are developed for simulation or analysis of the application.

TRL 3. *Analytical and experimental critical function and/or characteristic proof-of concept*: Proof of concept validation. Active Research and Development (R&D) is initiated with analytical and laboratory studies. Demonstration of technical feasibility using breadboard or brassboard implementations that are exercised with representative data.

TRL 4. *Component/subsystem validation in laboratory environment*: Standalone prototyping implementation and test. Integration of technology elements. Experiments with full-scale problems or data sets.

Fig. 2.10 The technology readiness level (TRL), barometer. Image NASA



- TRL 5. *System/subsystem/component validation in relevant environment*: Thorough testing of prototyping in representative environment. Basic technology elements integrated with reasonably realistic supporting elements. Prototyping implementations conform to target environment and interfaces.
- TRL 6. *System/subsystem model or prototyping demonstration in a relevant end-to-end environment (ground or space)*: Prototyping implementations on full-scale realistic problems. Partially integrated with existing systems. Limited documentation available. Engineering feasibility fully demonstrated in actual system application.
- TRL 7. *System prototyping demonstration in an operational environment (ground or space)*: System is at or near scale of the operational system, with most functions available for demonstration and test. Well integrated with collateral and ancillary systems. Limited user documentation available.
- TRL 8. *Actual system completed and 'mission qualified' through test and demonstration in an operational environment (ground or space)*: End of system development. Fully integrated with operational hardware and software systems. Most user documentation, training documentation, and maintenance documentation completed. All functionality tested in simulated and operational scenarios. Verification and Validation (V&V) completed.
- TRL 9. *Actual system 'mission proven' through successful mission operations (ground or space)*: Fully integrated with operational hardware/software systems. Actual system has been thoroughly demonstrated and tested in its operational environment. All documentation completed. Successful operational experience. Sustaining engineering support in place.

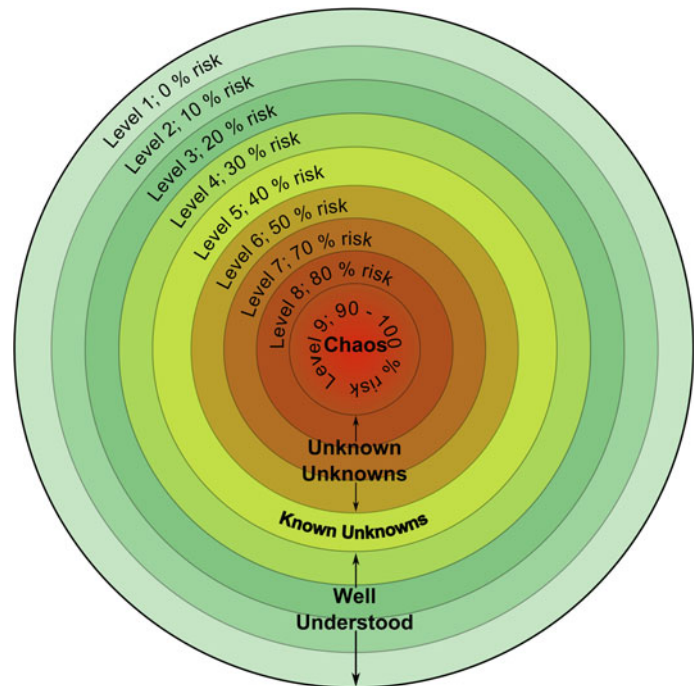
Understanding the TRL of a technology is critical to understanding the risk associated with that technology and, as such, an accurate assessment of a technology's TRL is a critical part of the mission analysis and design process. The application of TRLs to technology management will be discussed in [Chap. 21](#).

2.3.4 AD²: Advancement Degree of Difficulty

It was recognized within NASA when the TRL granularity was expanded from seven to nine levels in the mid-1990s that TRLs give an incomplete understanding of the technical concept, capability or product being assessed. As such in 1998 John Mankins, who had developed the increased TRL granularity proposed a 'Research and Development Degree of Difficulty', R&D³, system as "a measure of how much difficulty is expected to be encountered in the maturation of a particular technology" [1]. Within the R&D³ system five levels of difficulty were defined, giving the probability of success with 'normal' levels of research and development effort as between 20 and 99 %. Although the TRL concept is today widely used, the R&D³ system was never widely adopted or used.

Using the core principles of the R&D³ system, the 'Advancement Degree of Difficulty' (AD²), system was proposed in 2002 [2], focusing on the issues with the development and incorporation of new technologies into a space systems. As a result, the AD² system provides nine levels of risk, from 0 to 100 %, associated with the advancement of a technology from one TRL to the next, as shown in [Fig. 2.11](#). Only by combining TRL and AD², or some similar assessment, can a complete understanding be gained of the maturity and applicability of a technical concept, capability or product.

Fig. 2.11 Advancement degree of difficulty (AD^2), levels of risk.
Image Malcolm Macdonald



The AD^2 can be defined further as

- Level 1. Exists with no or only minor modifications being required. A single development approach is adequate.
- Level 2. Exists but requires major modifications. A single development approach is adequate.
- Level 3. Requires new development well within the experience base. A single development approach is adequate.
- Level 4. Requires new development but similarity to existing experience is sufficient to warrant comparison across the board. A single development approach can be taken with a high degree of confidence for success.
- Level 5. Requires new development but similarity to existing experience is sufficient to warrant comparison in all critical areas. Dual development approaches should be pursued to provide a high degree of confidence for success.
- Level 6. Requires new development but similarity to existing experience is sufficient to warrant comparison on only a subset of critical areas. Dual development approaches should be pursued in order to achieve a moderate degree of confidence for success. Desired performance can be achieved in subsequent block upgrades with a high degree of confidence.
- Level 7. Requires new development but similarity to existing experience is sufficient to warrant comparison in only a subset of critical areas. Multiple development routes must be pursued.

Level 8. Requires new development where similarity to existing experience base can be defined only in the broadest sense. Multiple development routes must be pursued.

Level 9. Requires new development outside of any existing experience base. No viable approaches exist that can be pursued with any degree of confidence. Basic research in key areas needed before feasible approaches can be defined.

2.3.5 ITAR: International Traffic in Arms Regulations

A further issue to consider in the availability of technology, especially for technologists outside the USA is the impact of the 1976 Arms Export Control Act of the US government, which gives the President of the United States the authority to control the import and export of defense articles and services. The provisions of this act are implemented within International Traffic in Arms Regulations, often termed simply ITAR. ITAR dictates that items on the United States Munitions List (USML) are export-restricted items. USML items are subject to change and re-interpretation. For example, following the February 1996 launch failure of the Long March-3B carrying Intelsat-708, which contained sophisticated communications and encryption technology, several parts of the spacecraft debris were never recovered by the satellite's American developers. This led to the suggestion that debris may have been recovered by the government of the People's Republic of China, with

Intelsat and the Clinton administration suffering domestic criticism for possibly allowing technology transfer to China. Following an investigation by the US Congress, in 2002 the United States Department of State charged Hughes Electronics and Boeing Satellite Systems with export control violations in relation to the failed launch of Intelsat-708 and the prior failed launch of the APSTAR-II satellite. As a result, space technology become subject to scrutiny within the ITAR framework.

The goal of ITAR is to limit arms proliferation, safeguard the national security of the US and further its government's foreign policy objectives. However, the selection of USML items can have significant adverse programmatic effects for space programs outside the USA, limiting, for example, launch vehicle options or even the end-customers access to the purchased system. As such, 'ITAR-free' components, sub-systems or even platforms are a major selling point for commercial components, sub-systems, systems and platforms in Europe and beyond.

The impact of ITAR was a reduction of the US share of the commercial spacecraft production market from 83 % in 1999, when the State Department took over the export regulation of spacecraft, to 50 % in 2008 [3]; moreover,

European manufacturers wherever possible avoid the use of ITAR (and hence US) components. In 2010, the US Congress requested an assessment of the risks of removing spacecraft and their components from the USML. The study, known as the 1,248 report, was completed in April 2012. In late 2012, the US Congress passed the fiscal 2013 defense authorization bill, which allows the president to remove commercial spacecraft and their components from the USML. It also allows him to decide which satellite technologies are the most important to protect while continuing to restricts export to China, Cuba, Iran, North Korea, Sudan, and Syria. The impact of this change, along with the effectiveness of its implementation, will take a number of years to assess.

References

1. Mankins, J.C., "Research & Development Degree of Difficulty (R&D³)", A White Paper, NASA Headquarters, March 1998.
2. Bilbro, J.W., Sackheim, R. L., "Managing a Technology Development Program", A White Paper, George C. Marshall Space Flight Center, May 2002.
3. "Earthbound", *The Economist* (U.S. Edition), pp. 66, August 21, 2008.

Henry B. Garrett

Just as spacecraft design teams are increasingly approaching the design and construction of a spacecraft as an integrated system, the overall environment and survivability of the spacecraft should be approached in a similar fashion. Typically perceived as either too expensive or design limiting, design for environmental survivability, whether it be from thermal, radiation, atomic oxygen, or spacecraft charging effects, is usually done strictly on an *ad hoc* basis. Unfortunately, ‘faster, better, cheaper’ (FBC) missions seldom consider anything much beyond thermal effects and, independently, radiation effects on selected parts. The basic requirements however, to significantly reduce the weight/size of a FBC mission and to make use of the latest commercial, off-the-shelf devices [with their often significantly lower radiation and Single Event Effects (SEE) tolerances] mandate that much greater thought be given to multiple uses of the spacecraft design to fulfill multiple environmental survivability functions. The objective of this chapter, after providing an introduction and overview of the space environment and its effects, is to detail the steps required for a systematic approach to space environment survivability that can be achieved with the least impact on the overall design process.

Fortunately, the concepts required to carry out a systems approach to environmental survivability currently exist. For example, both the Galileo Jupiter and the Cassini Saturn missions expended considerable effort in developing the methods necessary to design a thermal protection system that both provided meteoroid protection and limited spacecraft charging effects. In the case of Galileo, extensive effort was spent in developing an integrated radiation-resistance design for the Star Scanner—the designers picked a radiation resistant photomultiplier, substituted mirrors for lenses where

possible, and carefully placed additional shielding to provide robust protection (upwards of 10 g/cm²), to maximize the survivability of this system during Galileo’s passage through Jupiter’s inner radiation belts. Spot shielding, Error Detection and Correction (EDAC) software, hardening of selected components, Faraday cage shielding of the cabling, and similar techniques were all combined to provide ultra-reliable protection for the Galileo and Cassini systems. In the case of Cassini, attention was also paid to the way the vehicle was oriented in flight so as to limit meteoroid impacts. To a degree, radiation fluxes are also ‘oriented’—a factor that can be used to limit impacts on sensitive surfaces. On a case-by-case basis, good tools exist for providing specific types of environmental protection and that, in some instances, allow combining techniques. A well thought out survivable design considers all these components simultaneously.

The next generation of ‘microsats’, ‘cubesats’, or ‘sciencecraft’ will implicitly require a systematic approach to environmental protection if they are to realize meaningful levels of reliability within the size, mass, and power constraints of these concepts. The placement of parts, the selection of environmentally robust software (i.e., EDAC for SEEs), intelligent ‘on–off’ control of sensitive systems when the spacecraft is in a hazardous environment (many components are ‘harder’ when turned off), use of intrinsically hard circuit designs as opposed to softer circuit designs, redundancy, utilization of graceful degradation, multiple use of shielding (for thermal, radiation, spacecraft charging, atomic oxygen protection, etc.) are a few of the procedures to be considered. One example brings the point home: on the US Department of Defense Clementine spacecraft, officially called the Deep Space Program Science Experiment (DSPSE), the average shielding was ~100 mils (~2.5 mm) of aluminum. This implied that the solid-state recorder would be sensitive to approximately 1,000 Single Event Upsets (SEU) per day background due to protons. Indeed Clementine experienced an observable solar proton event

H. B. Garrett (✉)
Jet Propulsion Laboratory (JPL), California Institute
of Technology, Pasadena, CA, USA
e-mail: henry.b.garrett@jpl.nasa.gov

during its first month of operation. The Clementine solid-state recorder, however, did not see the event and averaged around only 70 SEUs per day over the mission. A careful review of the spacecraft design revealed that the majority of the solid-state recorder components were protected by at least 300 mils (7.6 mm) of shielding—not from the spacecraft but because the boards were closely packed inside their boxes and provided a significant amount of self-shielding. Designers, particularly in the early stages of a mission often fail to take account of such ‘intrinsic shielding’, leading to an erroneous concern for radiation effects.

A proper systematic design approach to environmental survivability requires: (1) a review of the primary environments and interaction(s) of concern and (2) a listing of the general design options for each concern. These options should be cross-correlated with the specific interactions to identify design options common to the different effects. The design is then iterated with changes in the design reflected in quantifiable metrics for each effect—for example, changing a thermal blanket design may change the meteoroid protection and may alter the radiation shielding and spacecraft mass. Changing the position of a star scanner might enhance its radiation protection or alter its thermal load. A systematic design approach needs to identify such ‘cross-correlations’. Ultimately, the goal of an analysis is to identify the minimum number of design procedures that can yield the maximum benefit for several different environmental effects.

The steps taken to limit a particular environment and its effects are typically well understood—Galileo is an example of how different protection methods can be played off against multiple effects. It is also clear that if mass and size are a premium and if environmentally ‘soft’ and advanced technology are synonymous, then the integrated approach is both necessary and a prerequisite if missions are to succeed in the future. Guidelines and methods for approaching the problem systematically are reviewed in this chapter. The objective is to provide an insight into initial integrated environmental survivability design and for establishing the reality of the potential benefits.

3.1 Procedure

The steps for identifying various integrated design trade-offs starts with the definition of the mission: its trajectory, instruments, and requirements. From these the relevant environments and interactions are defined. Based on the top-level interaction(s), the design trade-offs or options are then identified. These are then assessed in terms of relevant selection criteria (say, mass, cost, complexity, software impact, and so forth). The design trade space is optimized and a set of design solutions developed for project consideration. These steps are listed in Table 3.1.

Table 3.1 Integrated environmental design procedure

Step number	Step
One	Identify requirements based on trajectory, instruments, and unique mission constraints
Two	Rate the environments versus the interactions
Three	Identify the design trade-offs for the environments/interactions of highest concern
Four	Establish mass, cost, complexity criteria metrics for trade-offs
Five	Optimize combinations of design choices
Six	Evaluate resulting designs

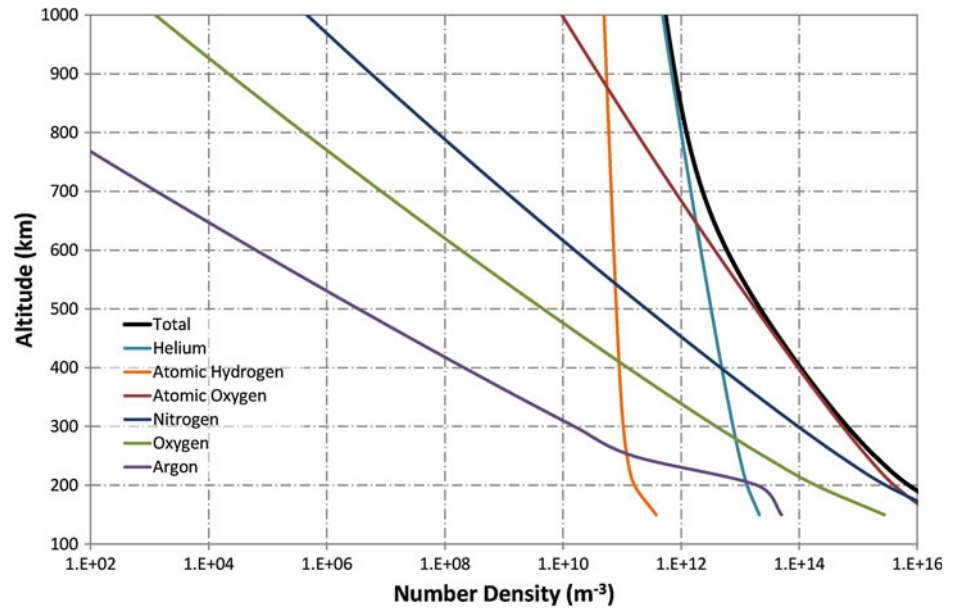
3.2 Environments

The space environment is far from benign in its effects on space systems. Given the growing complexity and consequent sensitivity of space systems, an understanding of the space environment and its interactions is the first step in mitigating these effects. Ten types of environment will be considered here. The first, the neutral atmosphere, is primarily responsible for drag, glow, and oxygen erosion. The next two environments, the magnetic and electric fields, are responsible for magnetic torques and induced electric fields. The UV/EUV radiation environment is not only responsible for the formation of planetary ionospheres but also for photoelectrons and long-term changes in material surface properties. The IR environment is a major driver of thermal effects. Four charged-particle environments are considered: the interplanetary environment, the plasmasphere/ionosphere (responsible for ram/wake effects and solar array arcing), the plasmashet (the primary region for spacecraft charging) and its low altitude extension the auroral zone, and the radiation belts. Although primarily referenced to the Earth, these environments each have their direct corollaries for the other planets as well. Finally, the solid-particle environment (synthetic space debris (unique to the Earth), interplanetary meteoroids, and surface dust) will be discussed (cometary particle clouds and planetary rings not also be considered). The intent is not to provide a detailed description of each environment (which are planet/orbit specific) but rather to provide an overview of their chief characteristics as they apply to environmental interactions. These characteristics are needed in defining the spacecraft effects for the purpose of design trade studies.

3.2.1 Neutral Atmosphere

Typically, the major environment at low altitudes around the planets (except Mercury) and Titan is the ambient neutral atmosphere. Atmospheric drag and ablation are

Fig. 3.1 Number density profiles for the Earth's atmosphere based on the US Standard Atmosphere, 1976 [2]



major concerns for this environment. In addition, typical orbital velocities relative to an atmosphere lead to impact energies of multiple eV's—high enough to induce chemical interactions such as oxygen erosion. Neutral particle densities for the Earth range from 10^{10} cm^{-3} at approximately 200 km altitude to 10^6 cm^{-3} or less at 1,000 km altitude, see Fig. 3.1. Atmospheric models that describe the density, composition, and temperature fall into three basic classes: static profiles, global analytic fits, and time-dependent simulations. Examples for the Earth are the US Standard Atmosphere profiles, the Jacchia and MSIS (mass spectrometer and incoherent scatter) analytic models, and various thermospheric global circulation models (TGCMs) [1]. For the Earth at least, there are a number of models. Similar types of models exist for the other planets and some of the moons. Static profiles in particular are readily available for most destinations, but aside from Mars, where several MSIS and TGCM models exist on-line, direct access to more complicated models for the other planets is typically limited. Static models for planets such as Venus and Mars are useful for reentry or atmospheric capture. The effects of the atmosphere on a spacecraft's trajectory, along with a more detailed discussion of atmospheric density models, can be found in the astrodynamics chapter (Chap. 4).

3.2.2 Electric and Magnetic Fields

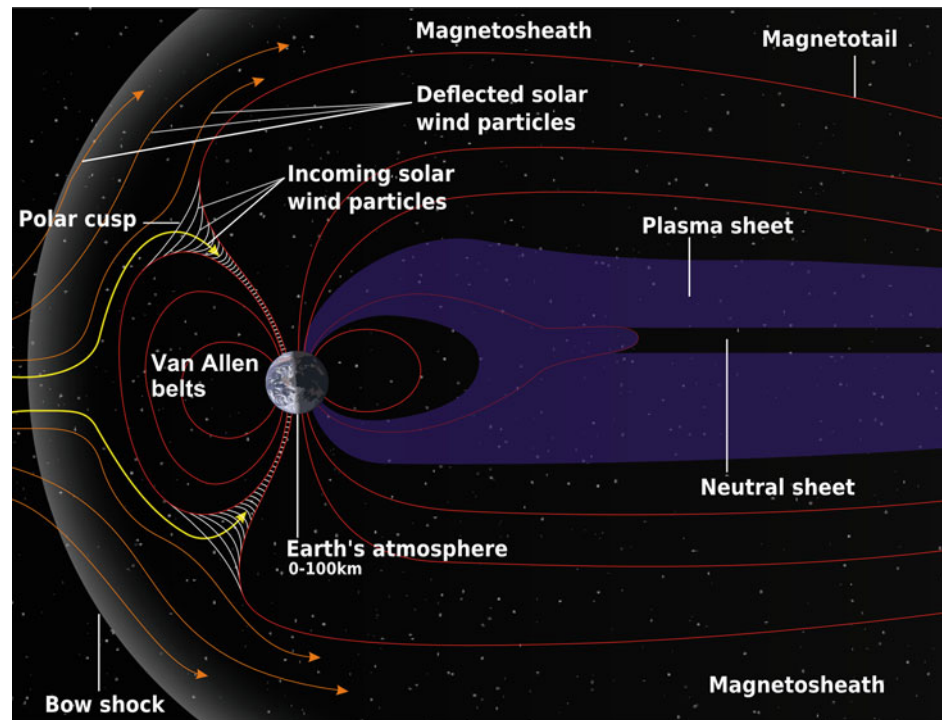
Electric and magnetic fields exist around most bodies in space. Magnetic fields range from tenths of a gauss near the Earth and $\sim 4\text{--}8$ gauss at Jupiter's surface down to a few gammas (nanotesla) in the solar wind. Note that one tesla is equal to 10^4 gauss. Ambient or induced electric fields (e.g.,

$\mathbf{V} \times \mathbf{B}$, see later) range from 0.3 V/m close to the Earth to as much as 60 V/m near Jupiter. For comparison, spacecraft surface charging potentials can reach ~ 20 kV at the Earth. There are detailed models of the magnetic fields of Earth, Jupiter and Saturn, from which the induced electric fields can be derived [3]. Mars, the Moon and Venus do not have significant magnetic fields, although strong local magnetic field anomalies have been identified at the Moon, while Mercury's field is apparently about 1 % as strong as Earth's but remains subject to further investigation by both the MESSENGER (Mercury Surface, Space ENvironment, GEochemistry and Ranging) and BepiColombo space missions. There are first-order models of the magnetic fields of Neptune and Uranus but these will need to be better developed in the future. Although simple magnetospheric models exist for all the planets, except for the Earth (Fig. 3.2) these need to be made more quantitative to determine actual magnetopause and magnetosheath crossings for instruments. Finally, computer codes capable of tracing out the field lines from the magnetic field models are needed for the radiation belt models (the latter typically require so-called 'B and L' coordinates) and are readily available. Although currently little information exists, the magnetic fields of comets and perhaps asteroids will also need to be defined during early planning of missions to these bodies.

3.2.3 Ultraviolet Radiation

Ultraviolet (UV) and extreme ultraviolet (EUV or XUV) radiation is important for spacecraft interactions as it can change the surface chemistry of materials and causes

Fig. 3.2 Profile of the Earth's magnetosphere showing the magnetic field lines and structure. Image NASA



photoelectron emission. The UV/EUV radiation is the continuum and line spectrum between roughly 10 and 4,000 Å. The solar flux/energy in this spectral range is between 10^7 and 10^{10} photons/(cm² s) below 1,000 Å and rises exponentially to 10^{16} photons/(cm² s) between 1,000 and 10,000 Å. Note that the Lyman-alpha line at 1,216 Å plays a major role in photoelectron emission. The shortest wavelengths, from 10 to 100 Å, are called X-rays. The solar spectrum at 1 au¹ is illustrated in Fig. 3.3a [4], while Fig. 3.3b presents the ASTM E490-00a(2006) Standard Solar Constant and Zero Air Mass (AM0) Solar Spectral Irradiance, which has an integrated power of 1,366.1 W m⁻². The ASTM E490 standard does not cover the complete solar spectrum but does extend from a wavelength of 119.5 μm to 1 m [5]; an ISO standard is also available, see ISO-21348. Note that the solar spectrum can also be represented simplistically as a black-body of effective temperature 5,781 K, whilst the Earth radiates as a black-body at 254 K. Models of the UV/EUV spectra and the atmospheric attenuation at the Earth and the planets of the UV/EUV are available if attenuation effects for sensors or photoemission are needed—models also exist for estimating spacecraft charging effects during eclipse passage.

¹ An astronomical unit is a unit of length defined as exactly 1.495 978 70691(6) × 10¹¹ m, approximately the average Earth–Sun distance, and is accepted for use with the *Système international d'unités*. The abbreviation is not capitalized as it is not named after a person; a.u. and ua are also used alongside the incorrect AU.

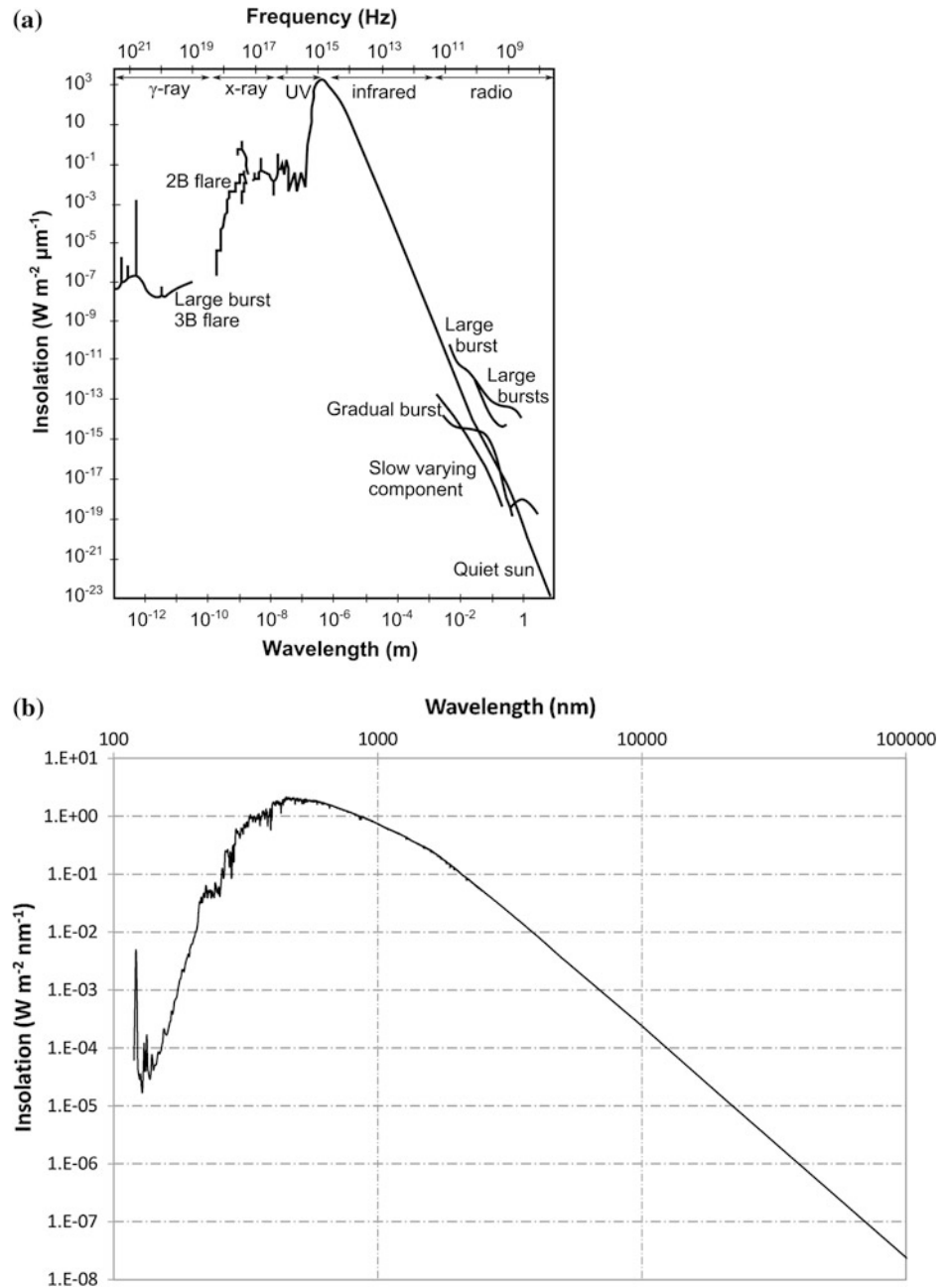
3.2.4 Infrared

The infrared (IR) spectrum is between roughly wavelengths of 0.7 and 7 μm and is dominated by the Sun (Fig. 3.3). Other sources of IR are reflected sunlight, atmospheric glow, radiation from planets, and even light from auroral displays. The IR environment is a major source of thermal effects on spacecraft. As in the case of the UV/EUV region, detailed spectra are readily available.

3.2.5 Solar Wind Plasma

The solar wind is a neutral plasma, primarily consisting of electrons, protons, and alpha particles, which flows approximately radially from the Sun at velocities ranging from 400 to 2,500 km/s. Since the Sun rotates in just over 27 days, as the solar wind expands outward the plasma drags the Sun's magnetic field lines out in an Archimedean spiral in the solar equatorial plane (Fig. 3.4). Densities (mean energies) range from around 50 particles cm⁻³ (~40 eV for ions; ~65 eV for electrons) near Mercury to 0.2 particles cm⁻³ (1 eV for ions; 10 eV for electrons) at Jupiter. Solar wind models are necessary for design purposes ranging from missions near the Sun to the outer solar system—a good example of such models is the NASA Marshall Space Flight Center (MSFC) L2-CPE statistical model [6]. Such models are used to estimate plasma interactions with spacecraft, large solar sails, comets, or asteroids, and for estimating effects on plasma sensors or

Fig. 3.3 a Electromagnetic flux at 1 au showing the frequency range from gamma (γ)-rays/X-rays through visible frequencies to IR and radio waves, reproduced from [4] and **b** the ASTM E490-00a(2006) Standard Zero Air Mass Solar Spectral Irradiance [5]



charging analyses. These plasmas induce spacecraft surface potentials of typically ~ 10 V—the highest reported surface potential in the solar wind being ~ 100 V. Plasma interaction models for estimating effects of the solar wind are available for general design purposes—examples are the Nascap-2 K [7] and various particle in a cell (PIC) codes.

3.2.6 Ionospheric Plasma

The ionized component of a planetary atmosphere, the ionosphere, is typically a comparatively dense, ‘cold’ plasma.

For the Earth, the composition varies from an O^+ dominated environment between ~ 200 and 500 km with a maximum density of about 10^6 cm^{-3} , to H^+ dominated above 1,000–1,200 km with densities from 10^5 cm^{-3} at 500 km, to 10^3 cm^{-3} or less above 2,000 km, see Fig. 3.5. All the planets (and most large moons) have ionospheres with compositions characteristic of their neutral atmospheres. In addition to spacecraft surface charging (typically of little concern compared to auroral-induced charging), ionospheres affect radio wave propagation and are important for their effects on spacecraft communications. Simple static profiles currently exist for all the planets and Titan.

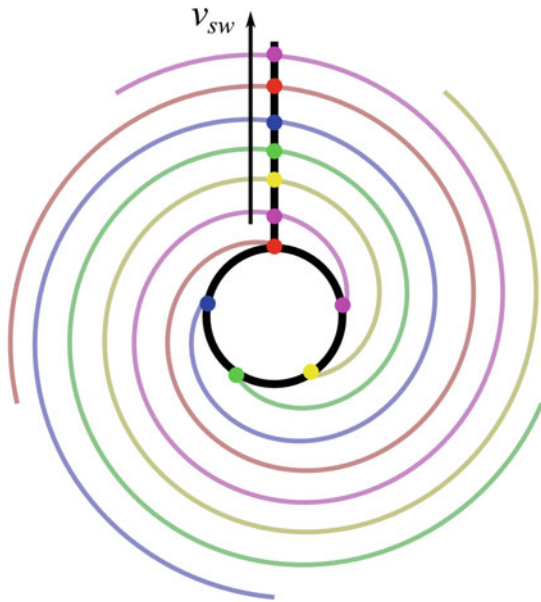


Fig. 3.4 View of the solar wind magnetic field lines showing how they are dragged out in an Archimedean spiral in the solar ecliptic plane as the Sun rotates. v_{sw} is the radial solar wind velocity vector

3.2.7 Aurora Plasma

Above the ionosphere and typically at the magnetic field boundary between high latitude, closed and quasi-closed magnetic field lines is a ‘hot’ plasma of substantially lower density than the ionosphere but much higher energy (the ‘plasma sheet’ region in Fig. 3.2). These particles (primarily electrons and protons) precipitate into the atmosphere generating bright arc structures called auroras. Near the Earth’s geostationary orbit (the equatorward extension of the auroral plasma), densities are on the order of $\sim 1 \text{ cm}^{-3}$ and mean energies of several tens of keV. This plasma can give rise to surface potentials of 20 kV or more. Auroras are regularly observed at Jupiter and Saturn and there are observations at Uranus and Neptune (Ganymede also has what appears to be auroras). As auroras pose a potential spacecraft charging threat, they need to be considered when evaluating a spacecraft’s charging mitigation system.

3.2.8 Trapped Radiation

Superimposed on the closed magnetic field lines of the ionosphere and auroral regimes are the high energy ($E > 100 \text{ keV}$) trapped electron and proton populations—the so-called van Allen belts. The important components are the electrons with energies between 100 keV to a few MeV and protons with energies from 100 keV to 100 MeV. Jupiter and the Earth have the most damaging radiation

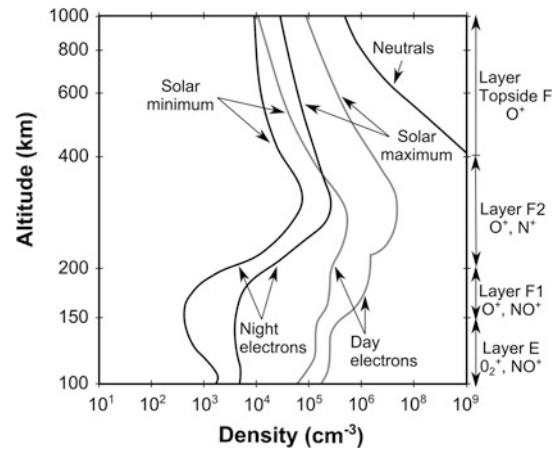


Fig. 3.5 Total ionization profile, with ionospheric layers, adapted from [4]

belts, though radiation belts exist at Saturn, Uranus, and Neptune. For terrestrial missions, the NASA AE8/AP8 radiation models have been the primary ones used, but the AE9/AP9 models will shortly supersede these. Jovian (GIRE) and Saturnian (SATRAD) radiation models are available from JPL. The Jovian model contains several ‘holes’; for example, a lack of a complete statistical understanding and proper modeling of time and pitch angle variations. Preliminary models have also been developed for Uranus and Neptune based on the Voyager flybys. The terrestrial and Jovian radiation belt contours for electrons and protons are illustrated in Fig. 3.6.

3.2.9 Galactic Cosmic Rays

The galactic cosmic ray (GCR) environment consists primarily of interplanetary protons and ionized heavy nuclei with energies from $\sim 1 \text{ MeV/nucleon}$ to higher than $\sim 100 \text{ GeV/nucleon}$. Electrons are also a constituent of GCR, but their measured intensities at energies above $\sim 10 \text{ MeV}$ are at least 1 order of magnitude smaller than the protons and are usually ignored. The principal element range of interest is from hydrogen to iron. Models of the GCR currently exist for interplanetary space and even for interstellar space (Fig. 3.7) as the Voyager spacecraft are currently crossing into the ‘pristine’ interstellar medium. Difficulties arise when modeling the detailed spectra for a given orbit within a magnetic field. Models for the Earth are available but the ability to model GCR transport at other planets is limited. For mission design purposes, it is normal to assume a ‘worst case’ environment (for example, ignoring magnetic shielding). Models of the in situ, trapped heavy ion environments at the Earth and Jupiter are also available for design purposes.

Fig. 3.6 Cross sections of the terrestrial (*top*) and Jovian (*bottom*) radiation belts. Fluxes are for 1 MeV electrons (*right*) and 10 MeV protons (*left*). Image I. Jun

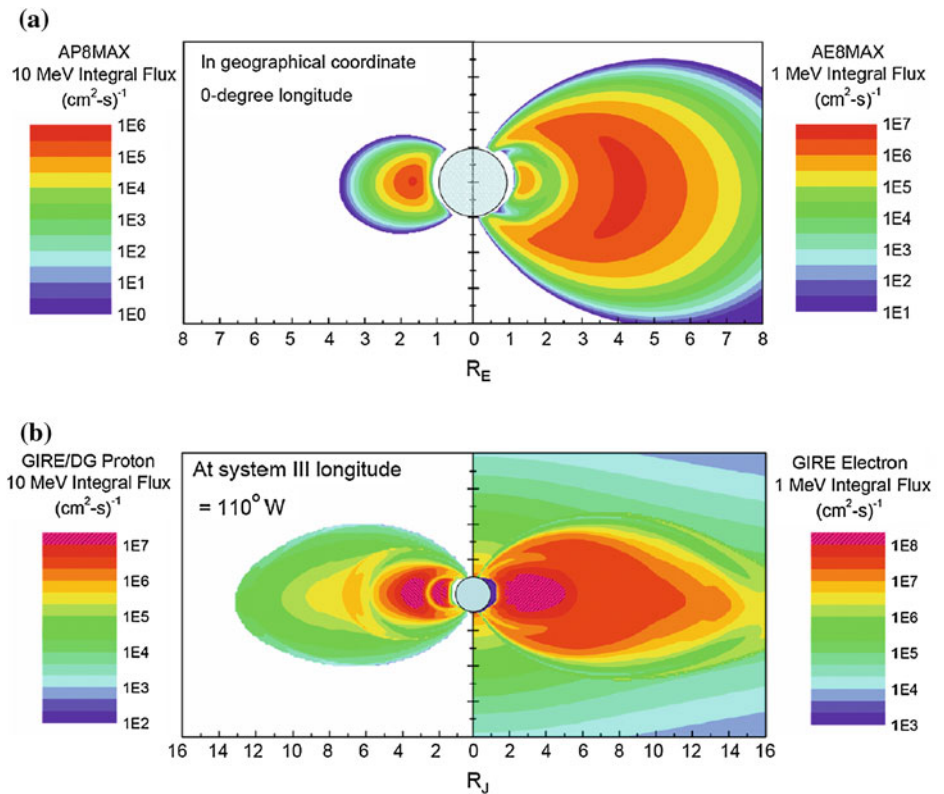
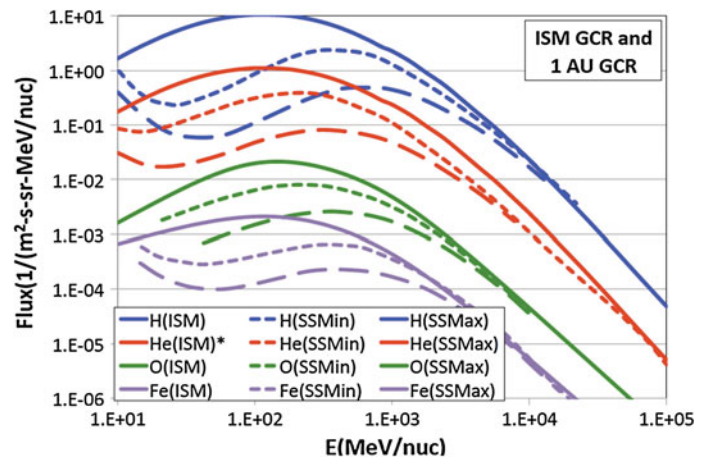


Fig. 3.7 Galactic cosmic rays (GCR) in the interstellar medium (ISM) and at 1 au. Shown are the proton (H), helium (He), oxygen (O), and iron (Fe) fluxes for the ISM (*top curve*), and solar minimum (SSMin-*middle*), and solar maximum (SSMax-*bottom*) conditions



3.2.10 Solar Proton Events

Hydrogen and heavy nuclei in the ~ 0.1 to ~ 100 MeV/nucleon energy range are ejected during a solar proton event (SPE) or, as it is also called, a solar energetic particle (SEP) event. Intensities are generally a few to several orders of magnitude larger than those of the GCR at these lower energies during these brief events (typically a few days or less in duration). The worst-case solar proton flux is approximately five orders of magnitude larger than the

GCR, but becomes 'softer' above ~ 100 MeV where the GCR begin to dominate the spectrum.

The energetic particles that make up these events are believed to come from two primary processes: acceleration at the surface of Sun in association with sunspots (so-called solar flares) or at the edge of a rapidly expanding coronal mass ejection (CME) in the solar wind. SPEs created by either process are, after trapped radiation, the major natural radiation of concern to spacecraft designers. Statistical models of the occurrence frequency of the largest events

have been developed that can be used to estimate doses for different mission lengths [8].

3.2.11 Meteoroids

Meteoroids are solid particles orbiting in interplanetary space (planetary ring material is a special case of ‘meteoroids’) and are believed to be either of cometary or asteroidal origin. The mass range is from 10^{-12} g dust grains to 10^{22} g for asteroids and comets (Fig. 3.8). Densities range from 0.5 g/cm^3 (fluffy ice) to between 3.5 g/cm^3 (stony) and 8.5 g/cm^3 (iron/nickel). Impact velocities range from 11 to 70 km/s (the latter particles are believed to be of interstellar origin) with mean values around 20 to 30 km/s. Currently, there are several models available, including the new MSFC Meteoroid Engineering Model (MEM) [9] and the older JPL-developed METEoroid Engineering Model (METEM) [10]. MEM incorporates the latest meteoroid data and is primarily intended for the 1 au environment. METEM provides interplanetary meteoroids from Mercury out to Saturn. The latter has modules for planetary focusing effects and planetary shielding. The METEM model is particularly useful for angular impact estimates and has seen wide use within the community. The database it uses is dated, however, and does not incorporate any of the new data that have become available since its debut.

3.2.12 Synthetic Debris

Space flight operations have led to an artificial shell of synthetic debris around the Earth. This shell of debris poses a greater threat than the natural meteoroid environment within 2,000 km of the Earth. Typical mass densities are 2.5 g/cm^3 and impact velocities are $\sim 10 \text{ km/s}$. The Orbital Debris Engineering Model ORDEM2000 by the Johnson Space Center (JSC) is perhaps the primary debris model and is available to download from the NASA Orbital Debris Program Office at JSC [11]. However, the ESA Meteoroid and Space Debris Terrestrial Environment Reference (MASTER) and Program for Radar and Observation Forecasting (PROOF) models are also widely available and are recommended by the ECSS standards. The ORDEM2000 model provides estimates of the near-Earth debris on a given date and unlike the ESA MASTER model, which is historical, includes spacecraft launch rates and impacts that can be used to project the future debris population. It should be noted however that ESA’s Debris Environment Long-Term Analysis (DELTA) tool can be used in conjunction with the MASTER tool to determine future debris trends. Representative debris fluxes are compared with the meteoroid environment in Fig. 3.9.

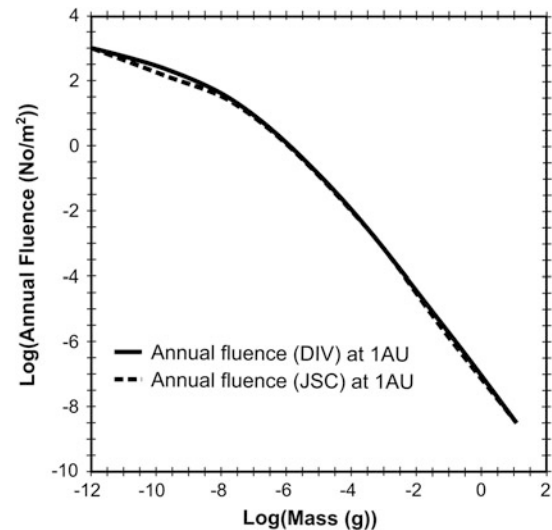


Fig. 3.8 Annual integral interplanetary meteoroid fluencies versus mass at 1 au for two standard meteoroid models

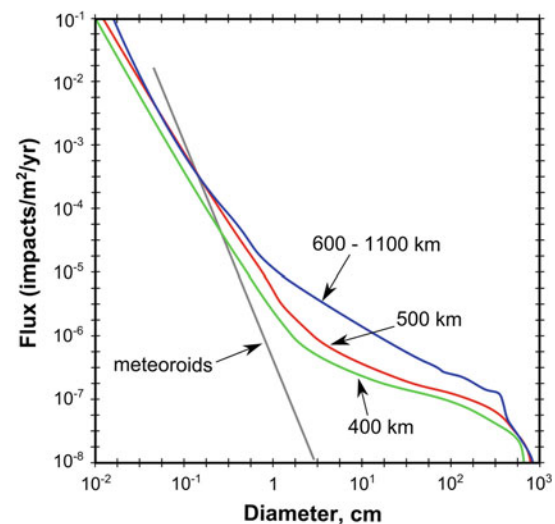


Fig. 3.9 Annual interplanetary meteoroid and space debris fluxes versus diameter at the Earth for various altitudes

3.2.13 Dust

An environment of increasing concern for which there are few models is dust. Significant dust environments have been observed at the Moon, Mars, and comets. Mars dust storms and dust devils and the very ‘sticky’ lunar dust that astronauts encountered are well-known problem environments. In the past, mission-unique models for specific comet missions have been developed but they are not well defined. Models of the in situ comet dust environment are important as this can seriously affect operations during flybys (hypervelocity impacts) or landings (contamination). ‘Dusty plasmas’ in this environment—dust that behaves as a

collection of charged particles—are another complication. This relatively new environment needs to be carefully considered because of the adhesion of the charged particles on surfaces.

3.3 Interactions

The anticipated sophistication and complexity of future space systems will greatly enhance their sensitivity to environmental interactions, and make what would otherwise have been second-order effects potentially critical problems for survivability. The purpose of this section is to review these interactions and relate them to possible areas of concern for the technologist.

Each category of interaction will be briefly defined in this section and examples provided of potential effects on a spacecraft and its subsystems. In defining these categories, it should be kept in mind that to some degree they overlap as several of the interactions are manifestations of a common underlying phenomenon (i.e., energy deposition or mechanical stress).

3.3.1 Cumulative Radiation Effects

Cumulative radiation effects depend on the type of particles, their energy, and their charge. A high-energy particle can transmute a material (change the atomic species and make the material radioactive), change its atomic structure (displacement damage), or produce free radicals, ions, and electron–hole pairs. Electronic parts and material characteristics thus slowly degrade with time due to these effects. A common measure of damage is total dose. Dose is the amount of energy deposited per unit mass of the absorbing material. An example is the total ionizing dose (TID) which is a measure of the energy deposited in a mass of material creating ionized charge pairs—typical units are 100 ergs/gm or 1 rad. Note that: the material has to be specified, e.g., ‘Si’ for silicon. For reference, commercial off-the-shelf (COTS) parts are typically ‘hard’ to sometimes as much as 10 Krads(Si), while space-qualified parts are typically ‘harder’ than 10 Krads(Si); ‘rad-hard’ parts are 100 Krad(Si) or higher and parts harder than 1 Mrad(Si) are ‘nuclear hardened’.

For engineering purposes, a dose versus depth curve is usually prepared for the design. In the early stages, this is done for a generic mass distribution—solid sphere, spherical shell, flat plate (or slab), two flat plates, etc. A set of calculations for the Clementine lunar missions is presented in Fig. 3.10. As shown in the figures, electrons are much more sensitive to the details of the shielding geometry than ions.

3.3.2 Single Event Upsets

The term single event effects (SEE) encompasses a variety of radiation-induced upsets in microelectronics. Of particular interest are single event upsets (SEU). SEUs are produced in an integrated circuit when a single charged particle passes through the circuit and causes a change in the state of a digital logic element leading to data loss or incorrect commands. As an energetic particle travels through a circuit element, it may deposit energy (producing ionization and a current pulse) sufficient to trigger the element (Fig. 3.11a). The energy loss is principally proportional to the square of the particles electrical charge, Z , but if nuclear interactions occur within the part, this rate can be substantially increased. Hence, more abundant low- Z ions deposit as much energy as less abundant high- Z ions. Figure 3.11b illustrates the actual effects on a Hubble Space Telescope CCD element—note the bright pixels that were ‘flipped’. The basic measure of energy transfer is linear energy transfer (LET) typically given in $\text{MeV cm}^2/\text{mg}$, which is the energy lost by the particle to the material per unit path length (MeV/cm) divided by the density of the material (mg/cm^3). Hence, multiplying LET by the density of the material being impacted gives the energy deposited per unit length in the material. The SEU rate for each circuit element needs to be evaluated for all the sensitive devices associated with a subsystem on the spacecraft. Figure 3.12 presents representative SPE and GCR fluxes versus LET for varying levels of shielding—shielding has a large effect on SPEs but not much on GCRs; parts with LETs above ~ 30 must be selected to significantly reduce GCR rates.

3.3.3 Latchup

Another form of an SEE is latchup. The passage of an energetic particle through a sensitive device can sometimes create a transient short circuit or current path. In the case of latchup, this can turn on a parasitic silicon controlled rectifier (SCR) resulting in either a loss of circuit function or thermal runaway from the excessive current. The latter can cause permanent damage. If detected early, both can be mitigated by powering down the device. Given the possibility of permanent damage by this effect, it needs to be evaluated for all potentially sensitive integrated circuits.

3.3.4 Surface Charging/Wakes

Surfaces immersed in a space plasma will charge to a potential relative to the plasma. In sunlight, this is typically a few volts positive due to photoelectron currents. In shadow, to first order, the potential is proportional to the

Fig. 3.10 Representative dose/depth curves for the Clementine lunar mission showing the differences between different shielding geometries for protons (a) and electrons (b). Note that 1 mil = 1/1,000th of an International Inch, which is exactly 25.4 mm

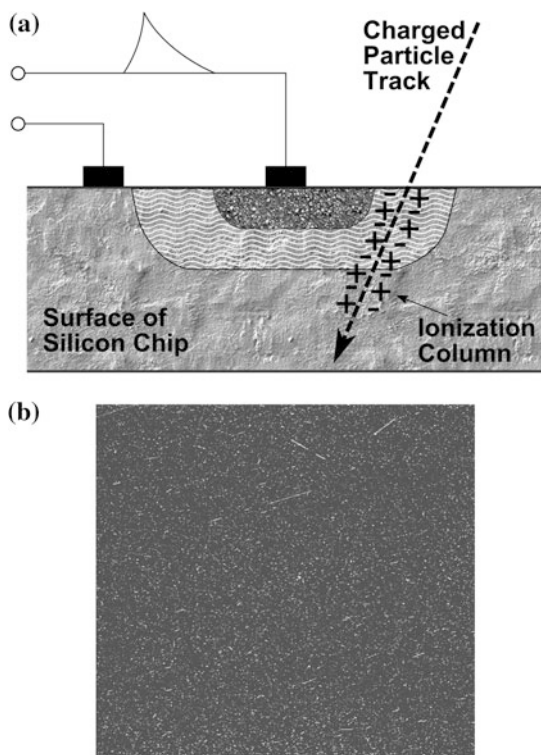
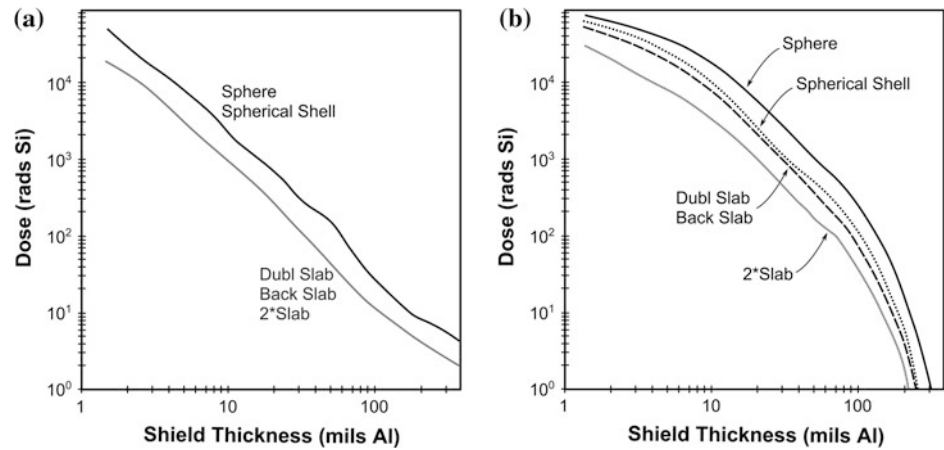


Fig. 3.11 Examples of SEU effects. **a** Schematic illustrates the process of charge deposition in a microcircuit element that leads to a bit flip. **b** CCD image from the Hubble Space Telescope as it passed through the South Atlantic Anomaly showing the effects of SEU events on the CCD pixels

ambient electron temperature (Fig. 3.13a) and current. These potentials can be over 10–20 kV (negative relative to the space plasma ground) and can produce differential potentials over 1,000 V between electrically isolated surfaces. Representative surface potentials in eclipse for the Earth are presented in Fig. 3.13b. As arcing may occur between charged surfaces with potential differences as little as 200 V, this can be a serious environmental concern.

In conjunction with surface charging, a plasmashield is often created, the scale of which is characterized by the Debye length, denoted as either λ_D or L_D . The formation of this sheath can be understood by considering the effect of placing a surface with no initial net charge into a plasma. At first, more electrons than ions will strike the surface due to the higher thermal speed of electrons; high-energy electrons can penetrate several millimeters, charging internal dielectrics, while lower energy electrons and ions deposit charge on the surface. This electron/ion strike rate imbalance causes the surface to charge negatively until the charge is sufficient to repel further electrons and to attract ions. At equilibrium, the electron/ion currents to the surface balance. A region is formed above the surface within which the positive ions outnumber the electrons, shielding the negative surface potential. Outside the structure, that is the plasmashield, the ambient plasma does not see the net negative potential.

In the ionosphere the Debye length is typically less than approximately 10 cm, however in the magnetosphere it can be from 0.1 to 1 km. If the Debye length is smaller than the characteristic length of the spacecraft the plasmashield will provide a conductive path between different parts of the spacecraft, keeping the potential relatively even. However, large potential differences can still occur in the wake region, where the Debye length can be locally large. If the Debye length is larger than the characteristic length of the spacecraft then large potential differences can develop on electrically isolated spacecraft surfaces causing potentially damaging arc discharges that can be a serious concern for solar cell systems, where exposing a semiconductor to sunlight is a required condition of operation. Debye shielding can have a serious effect on particle and field detector instruments, as the presence of the spacecraft alters the very field that the instruments are attempting to measure. As such, instruments are typically placed on electrically isolated structures that extend beyond the Debye shield of the spacecraft and can be biased relative to the spacecraft ground.

Fig. 3.12 Annual GCR and solar proton event fluxes for various shielding thicknesses. Image J. M. Ratliff

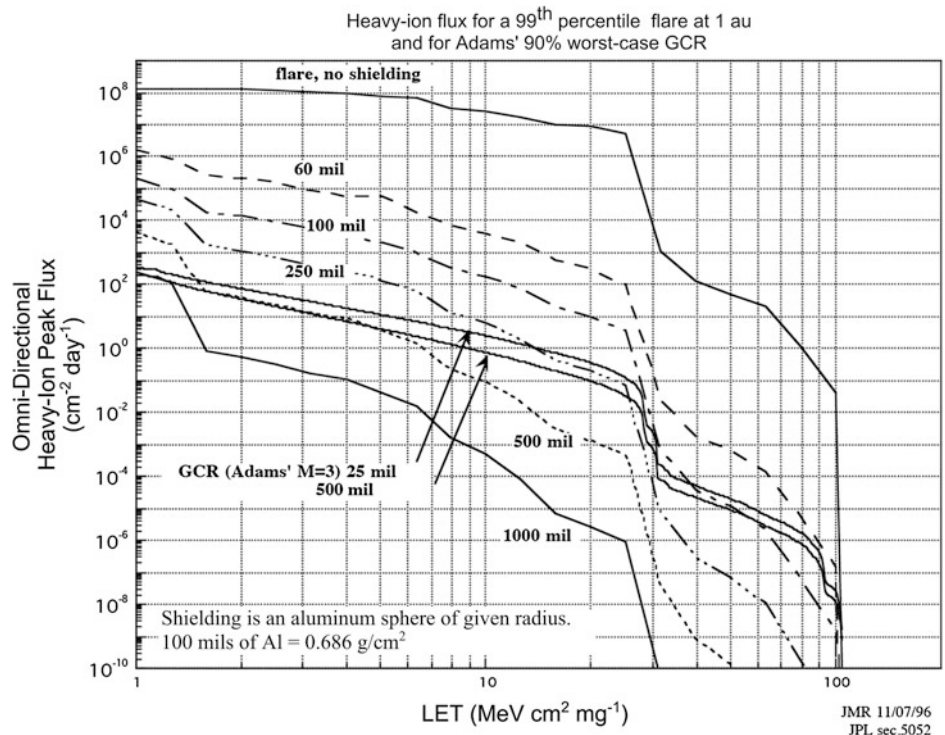
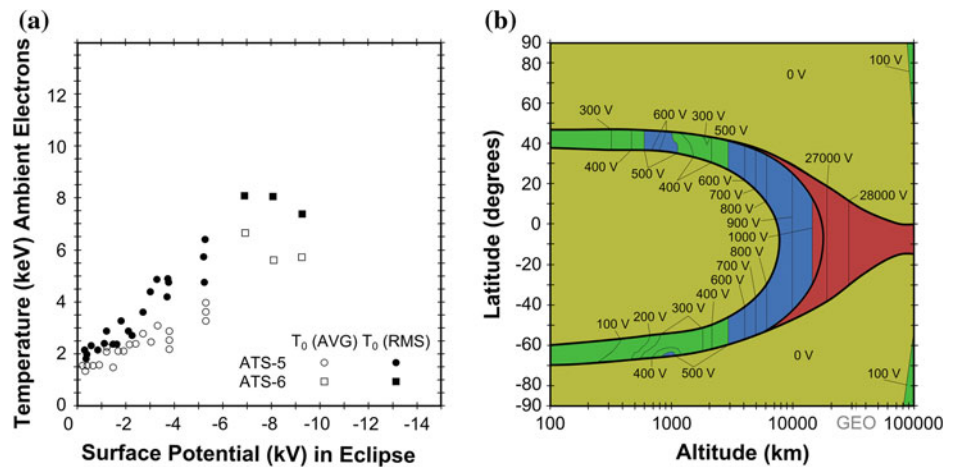


Fig. 3.13 Surface charging effects for the Earth. **a** Comparison of observed surface potentials in solar eclipse at geosynchronous orbit for the ATS-5 and ATS-6 spacecraft versus plasma temperature. **b** Estimates of the surface potential in eclipse versus position for the midnight meridian



Though normally of little concern, in moderate to dense plasmas like the ionosphere the vehicle's velocity relative to the plasma can produce a charged wake structure around the vehicle (and sensors) altering the currents and electromagnetic fields around it. Ionospheric ions typically have thermal velocities lower than the orbital velocity; as such, the motion of the structure causes the plasma density to build up in the ram direction, with a consequently low-density region occurring in the structure's wake. Density deviations can be several orders of magnitude from the ambient and the Debye length in wake can be locally very large, which tends to be negatively charged as plasma electrons typically have a greater velocity than ions. In addition to distortions in particle and field measurements, the dense plasma can

also lead to enhanced power loss (positive surfaces that draw electrons) to arcing (negatively charged surfaces). The International Space Station (ISS) flies plasma contactors to minimize these effects.

3.3.5 Internal Charging

In addition to surface charging, internal electrostatic charging/discharging (IESD) (also called buried charging) is a very real concern—particularly at the Earth (Fig. 3.14) and Jupiter. High energy electrons (100 keV or higher) can easily penetrate spacecraft surfaces and deposit charge on or in internal surfaces, but protons of the same energy are stopped

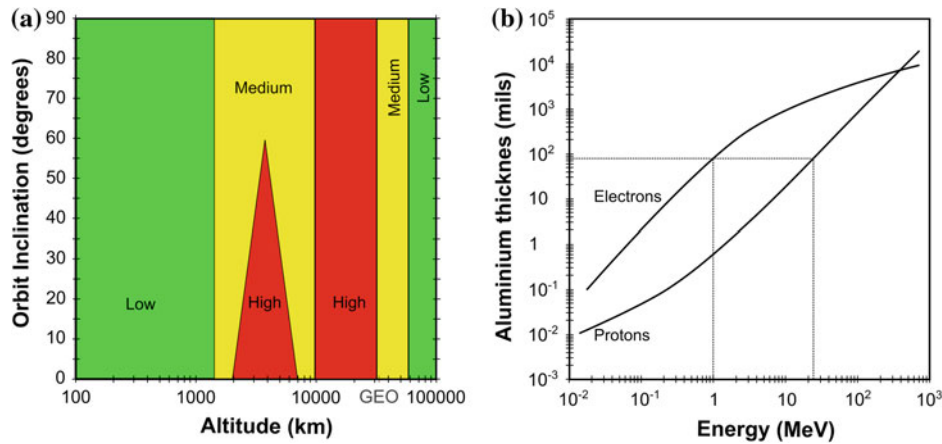


Fig. 3.14 Internal charging effects for the Earth. *Left* provides estimates of regions of IESD concern for circular orbits. *Right* gives the mean penetration depth of electrons and protons in aluminum. As illustrated, a 1 MeV electron penetrates as deeply as a proton of over

at the surface (Fig. 3.14)—the resulting differential charging can lead to arcing inside the normal Faraday cage of the spacecraft. At least 2.5 mm of aluminum shielding is typically needed in the Earth’s environment to prevent this effect. Methods for estimating the effects of internal and surface charging and mitigating their effects are presented in [12].

3.3.6 Power Loss

Although normally of limited concern, as power systems (particularly solar arrays and electrodynamic tethers) approach 200 V or more in operating voltage, positive potential surfaces can experience parasitic power losses. Any exposed positive surfaces (perhaps pin holes in insulation such as produced by micrometeoroid impacts) will attract electrons—for potentials of approximately 200 V or higher, the electrons will receive sufficient energy to generate dense clouds of secondary electrons. This plasma cloud effectively defeats the insulation and results in high ambient electron currents and power loss over positively charged surfaces.

3.3.7 Lorentz Effect

A conducting body crossing a magnetic field will experience an induced electric field proportional to the cross (or vector) product of the instantaneous velocity, \mathbf{V} , and the magnetic field, \mathbf{B} , that is $(\mathbf{V} \times \mathbf{B})$ —the Lorentz effect. In low Earth orbit, this can be as high as 0.3 V/m. Much higher values are experienced at Jupiter—approximately 60 V/m over the polar caps for the Juno mission. At the Earth, induced voltages of 10 V have been seen on the Shuttle and over

20 MeV (as there are many times fewer protons at the higher energy, negative charge builds up leading to IESD). Note that 1 mil = 1/1,000th of the International Inch, which is exactly 25.4 mm

100 V on the ISS. For a 100 km long tether (possible with today’s technology), potentials of approximately 10,000 V could be generated and the tether used as a power source (in return for a decrease in altitude). Given the varying nature of the potentials across a structure as it rotates, $\mathbf{V} \times \mathbf{B}$ potentials can be very annoying for some spacecraft.

3.3.8 Surface Damages

Arc crazing/blow-off, sputtering, ablation due to the neutral atmosphere, EUV-induced chemical changes, radiation damage (especially for Teflon), and other effects can seriously damage exposed spacecraft surfaces. Oxygen erosion in low Earth orbit has been found to be a serious problem for many organic compounds (Kapton in particular) and a few metals (silver and osmium). Even a few days spent in low Earth orbit can seriously damage some types of surface and, in the case of the Long Duration Exposure Facility (LDEF), entire surface samples were found to disappear after a few years of exposure. Finally, micrometeoroid impacts can fracture solar array cover glasses, penetrate cabling, and similarly degrade surfaces. Such degradation will lead to long-term decay in surface properties and must be considered in the selection of surface materials and appropriate coatings or shielding.

3.3.9 Contamination

Outgassing, thruster firings, gas leaks, water dumps, erosion of surfaces, flaking of paints, and long term curing of epoxies can all contribute to the contamination environment around a spacecraft. Charging can lead to enhanced

deposition rates on some surfaces while EUV and radiation can alter the chemical effects of the contamination. From changes in alpha/epsilon to the glint of small contaminate particles in the field of view of a sensor or the degradation of optical transmission properties, contamination is a serious problem. Water in particular is a pervasive and potentially highly damaging contaminant (nitrogen purges and expensive ground handling techniques are consequences). Control and limitation of such contamination effects is an important factor that needs to be included in a survivable design.

3.3.10 Atmospheric Glow

Although it has only been detected at the Earth so far, serious optical contamination in the form of a visible glow on surfaces facing into the spacecraft velocity vector has been observed for orbits of 800 km altitude or lower. Figure 3.15 shows an example of this glow along the vertical stabilizer and Orbital Maneuvering System (OMS) pods for the Shuttle. The phenomenon may result from the interaction of atomic oxygen with spacecraft surfaces, as the glow intensity appears to vary with the atomic oxygen density. The interaction generates optical emissions, primarily in the orange range of the spectrum (apparently consistent with the emission spectrum of NO_2) that can contaminate sensitive IR sensors. As the glow appears to come primarily from surfaces in the ram direction and to be enhanced during thruster firings, careful placement of optical sensors and timing of thruster firings may need to be considered in the mission design.

3.3.11 Particle Impacts

Hypervelocity impacts from a few km/s and up between meteoroids or synthetic space debris and spacecraft can be devastating. Interplanetary meteoroid impact velocities average between 20 and 30 km/s (impact velocities as high as 500 km/s, however, may occur near the Sun during a close perihelion passage) whereas space debris impacts are typically 10 km/s. At the Earth, particles with velocities of approximately 70 km/s or greater are believed to be of interstellar origin. Effects range from pitting to complete penetration of walls or even total destruction of a spacecraft. Wiring and pressure vessels (for example, crew quarters or fuel tanks) are particularly sensitive to these effects. Meteoroid shielding is thus a very important consideration for many missions—particularly those to the outer planets where even small pits in the engine nozzles or fuel tanks could lead to catastrophic failures. Externally exposed long cable runs and wire antennas are of particular concern as these are usually thin but very long leading to large areas (greatly increasing their likelihood of getting hit) which could be severed by

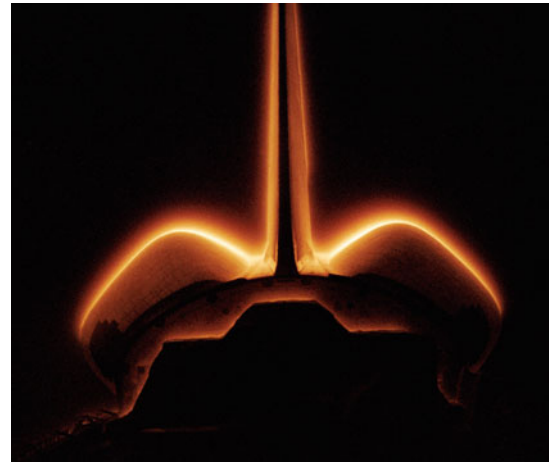


Fig. 3.15 Space Shuttle Columbia during a night pass on STS-62, March 1994. Image documents the glow phenomenon surrounding the vertical stabilizer and the Orbital Maneuvering System (OMS) pods of the spacecraft; NASA Photo ID: STS062-42-026. *Image NASA*

relatively small particles. Figure 3.17 illustrates the effects of a hypervelocity particle impact on a plate—the particle comes in from the left and exits on the right. Note how the debris cloud expands in a roughly spherical shape (Fig. 3.16).

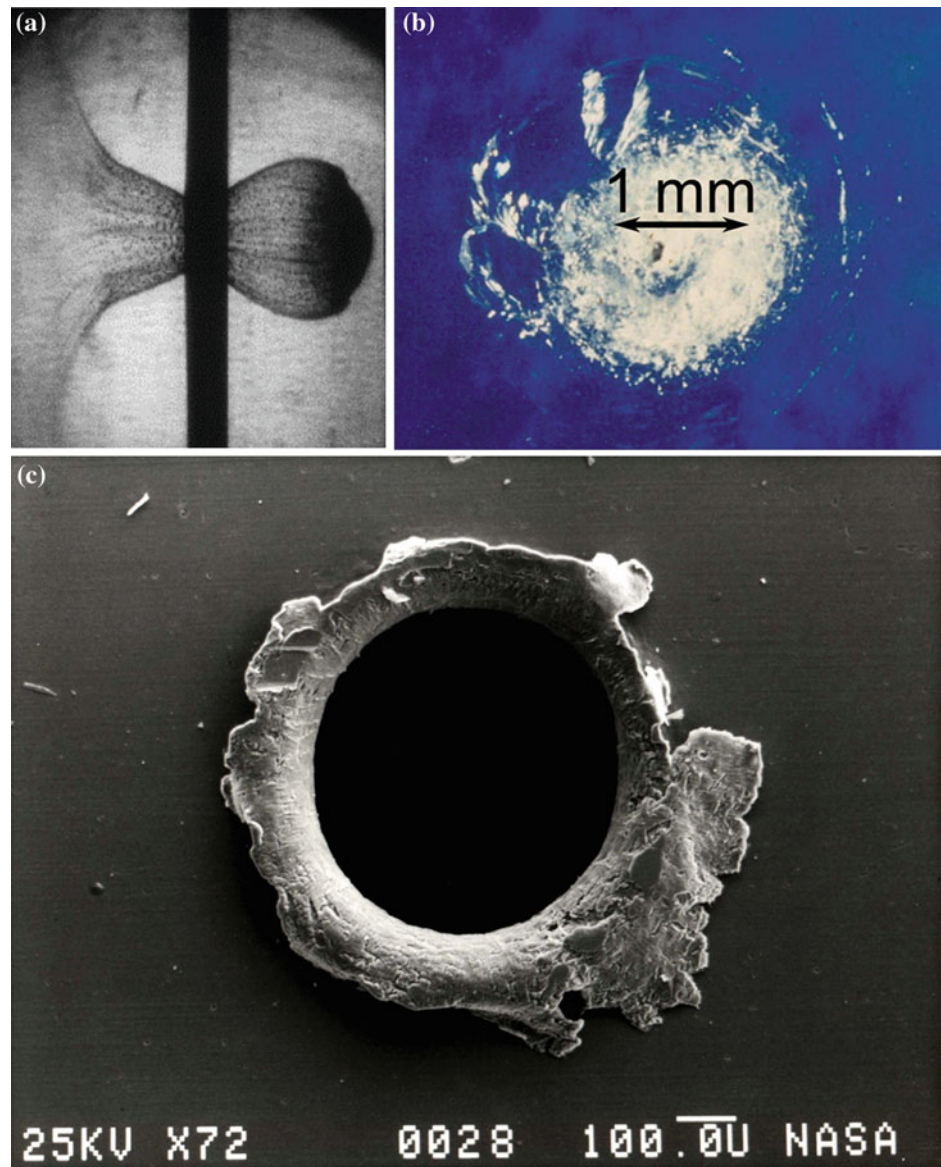
3.3.12 Torques

The effects of small forces on the stability of spacecraft are well known to cause degradation in pointing accuracy and mechanical deformation. Thermal effects (for example, expansion/contraction of booms), light pressure, gravity gradients, atmospheric drag, meteoroid impacts, and magnetic torques can all cause instabilities. Even arc discharges can impart a measurable impulse. The potential torques on a spacecraft associated with 500 km/s impacts near the Sun may be particularly critical for this class of missions. The sensitivity of the spacecraft to such torques needs to be evaluated for each class.

3.3.13 Thermal

Thermal effects (specifically, the effects of varying temperature and thermal radiation on components) are probably the most important environmental concern for spacecraft. Of particular concern are issues associated with the thermal protection system, as it is often intimately involved with the design of the exterior surfaces of the spacecraft. To reduce mass, design efforts should concentrate on integrating the thermal blanket layout with the meteoroid and spacecraft charging mitigation systems. Typical areas of concern are the conductivity of the thermal blankets, nuclear battery [i.e., radioisotope power source (RPS)] or nuclear reactor

Fig. 3.16 Effect of hypervelocity impacts; **a** the cloud of particles produced by a hypervelocity impact, **b** a window pit from orbital debris on the Space Shuttle Challenger during STS-7 and **c** a view of an orbital debris hole made in the panel of the Solar Maximum Mission, SolarMax, satellite



placement, optical surface reflectors (OSR), thermal sensors (which can pick up stray EM pulses), and, for missions close to the Sun, heat shields. As white paints and Kapton-based thermal blankets, the most common thermal control solutions, are typically non-conductive and sources of arc discharges, there can be complex trade-offs when carrying out a survivability design.

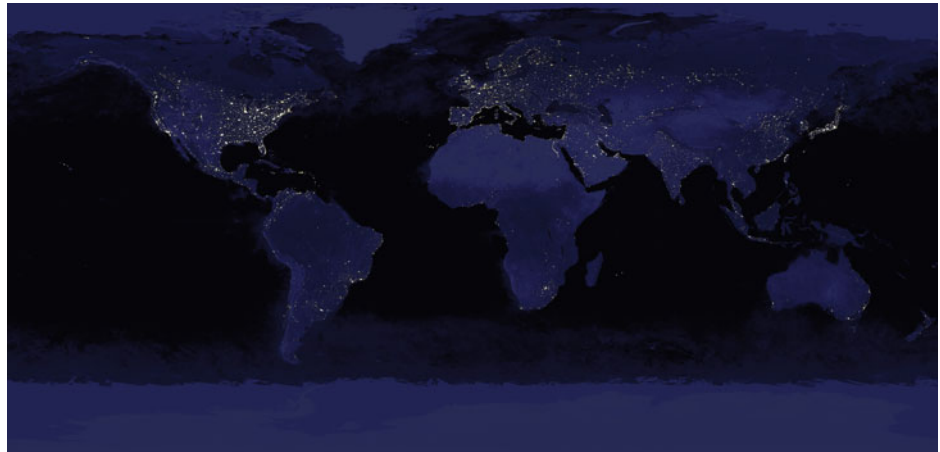
3.3.14 Other Effects

Density fluctuations in the ionosphere lead to enhancements and depletions in the electron density along a radio frequency (RF) propagation path. These processes can distort the phase and amplitude of a signal to and from a spacecraft. In addition to the Earth, typical ionospheric properties

and actual measurements are available for several of the planets and the Sun (radio occultation measurements being a common source). Auroras, the galaxy, and the Sun are all natural sources of background RF that can further hamper communications in space. Except in special cases, they are considered to be of secondary importance to typical electromagnetic compatibility (EMC) sources on the spacecraft itself and are ignored for design purposes.

The ambient environment has numerous sources of stray light. Besides the Sun itself which causes significant noise across the entire electromagnetic spectrum, stars, starlight, gegenschein (German for 'counter shine'; a faint brightening of the night sky in the region of the antisolar point), the zodiacal light (a faint, roughly triangular, whitish glow seen in the night sky that appears to extend up from the vicinity of the Sun along the ecliptic or zodiac), atmospheric glow, the

Fig. 3.17 Image of Earth's city lights created with data from the Defense Meteorological Satellite Program (DMSP), Operational Linescan System (OLS). Originally designed to view clouds by moonlight, the OLS was used to map the locations of permanent lights on the Earth's surface. *Image NASA-GSFC*



auroras, the equatorial electrojet, the polar cap aurora, and moonlight all contribute background light in the UV, EUV, and IR. Lately, at the Earth even city lights and oil flares have become a serious concern in making ground observations from space; see Fig. 3.17 and other representative images from the Defense Meteorological Satellite Program (DMSP) which have been particularly useful in bringing this problem to the attention of the public. These effects are considered to be of secondary importance in an integrated design (Fig. 3.18).

Rapidly oscillating fields on surfaces can cause serious interactions with space plasma. In particular, depending on the electron density, the mechanical spacing of the elements, and other factors, a resonance phenomenon called multipacting can be induced between the surfaces. Briefly, electrons accelerated into the surface by the time-varying positive component of the field can generate secondaries. These secondaries in turn (if the spacing and timing are correct) can generate more secondaries when they impact, causing a plasma avalanche. The cloud of electrons can lead to significant losses in signal strength and drain power from a transmitter. Whereas power loss is primarily a direct current (DC) process, this is an alternating current (AC) effect. The simulation of this phenomenon, unlike many other environmental effects, is fairly straightforward with the ambient environment playing only a secondary role once the process is initiated as the secondary electrons created quickly outnumber the ambients. Even so, systems must take this effect into account and be tested under ionospheric plasma conditions and, if possible, under the appropriate neutral atmosphere conditions before flight.

3.3.15 Interactions Versus Environments Trade Matrix

The matrix in Table 3.2 compares the key design environments for spacecraft with the critical interactions. For every

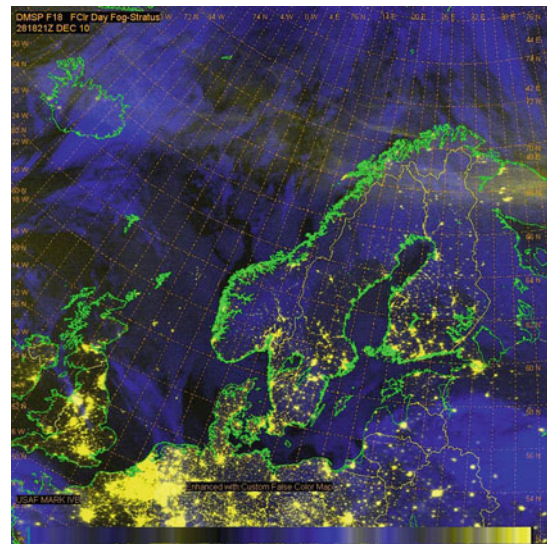


Fig. 3.18 A Defense Meteorological Satellite Program (DMSP) image from December 28, 2010 showing auroras over the polar regions north of Scandinavia outshone by the city lights of Europe. The image uses both nighttime visual and infrared imagery from the DMSP satellites F17 & F18. *Image US Navy Fleet Numerical Meteorology and Oceanography Center*

mission the technologist needs to identify the relevant mission environments and then determine the specific interactions of concern for that mission and its different phases. This has been done for three representative cases: a mission to Europa (E or e), an outer solar system mission to the dwarf planet Pluto (P or p), and a solar probe-like mission (S or s). Also indicated is whether the interaction is a major (capital letter) or minor (lower case letter) concern. Such assessments are very dependent on the specific mission but are helpful in identifying the critical interaction concerns. As such, this should be the first step in developing an integrated, survivable design.

3.4 Design Options

In this section the design trade space, namely the design options available to the reliability engineer for mitigating specific interactions will be described. Examples range from shielding for meteoroid impacts and blankets for thermal control to the careful selection of conductive materials for spacecraft charging and radiation protection. In particular, Tables 3.3 and 3.4 list some of the major design options for mitigating space environment effects. Representative mitigation options are discussed in more depth in the following sections.

3.4.1 Shielding

Probably the best-known environmental mitigation method is shielding. A number of shielding methods exist for mitigating specific interactions. For radiation and hypervelocity impacts in particular, shielding has proven to be a primary means of limiting their effects. Indeed, as the major objective is typically to place physical mass between the impactor (atomic or particulate) and the target, the methods for mitigating both are similar. The first step is typically to account for the intrinsic spacecraft body shielding. Where necessary, sensitive elements such as radiation-soft electronic components can be protected by bulk shielding, perhaps by encasing assemblies or spacecraft subsystems within a box of high-Z materials—referred to as a doghouse or vault. Or, for the case of hypervelocity impacts, by a specially designed, multi-layer meteoroid Whipple shield for crew quarters or fuel tanks. Special shielding plates can be placed between assemblies or slices in a stack. A specific technique for particularly radiation-soft parts is a so-called spot shield—this can be effective in conserving shielding mass. For other interactions there are similar techniques. In the case of electromagnetic compatibility (EMC) and electromagnetic interference (EMI), often termed EMC/EMI, an electrically conductive Faraday cage is usually employed. Baffles and deflectors are used to limit thruster contamination while in-flight removable covers may be used for science instruments.

3.4.2 Positioning

Closely tied to the use of shielding is positioning—the placement or orientation of critical components or surfaces to minimize the effects of the environment. Examples are the placement of radiation sensitive devices as far as possible from a nuclear power source or the orientation of a surface relative to the velocity vector to minimize the meteoroid fluence. Each of these forms of positioning can yield significant reductions in specific effects and may have

little or no impact on a system design. Examples are listed in Table 3.3.

3.4.3 Material Properties

Proper material selection for a specific environment can significantly improve the lifetime and reliability of a space system. Careful selection of materials can both limit and prevent many types of environmental interaction. As discussed earlier, there may be complex trade-offs between conductivity requirements, thermal paints, and radiation sensitivity. The steps required to identify the appropriate materials for mitigating interactions are listed in Table 3.4.

3.4.4 Electronic Parts Selection

Electronic parts are a critical component of a spacecraft design. Parts engineers need to evaluate a range of parameters in identifying the appropriate components for a specific mission. A parts engineer must trade cost and availability versus the class of a part—for example, radiation-hard, space-qualified (Class S or Class B) parts versus commercial parts. Note that failure rates of Class S parts are generally about a quarter of the rate for Class B parts [13]. The relationship between the classes is detailed in Table 3.5 [13]. The parts engineer must trade the known greater cost and limited availability of radiation-hard parts against the costs of radiation testing or the advantages of more capable commercial (non-radiation-hard) parts. Note that typically redundancy is not a justification for using a lower class part as it provides a lower reliability payoff at the point where it is needed. Maverick parts, production flaws, and other uncertainties, however, justify redundancy for critical circuits in high-reliability, long-life applications to protect against random failures [13]. For long-life, the use of high-reliability hardware, Class S parts, and redundancy in critical applications can provide an optimum and cost-effective approach. Parts also need to be evaluated for their electrostatic discharge (ESD), sensitivity. Increasingly, checks must also be made for counterfeits and, in the US and Canada, technologists should review parts lists for Government-Industry Data Exchange Program (GIDEP) alerts.²

3.4.5 Circuit/System Design

Careful circuit and system designs can be used to limit and mitigate the effects of radiation and thermal effects—there are for example circuits that are designed to compensate as

² See <http://www.gidep.org/>.

Table 3.2 Key environments and interactions matrix for three representative missions to Europa (E, e), Pluto (P, p) and a Solar Probe (S, s)

Environments	Interaction												
	Cumulative radiation effects	Single event upsets	Latch-up	Surface charging/wakes	Internal charging	Power loss	$V \times B$	Surface damage	Contamination	Glow	Particle impacts	Torques	Thermal
Neutral atmosphere													
E, B fields				ES			e						E
Ultraviolet radiation				S				S	S				
Infrared radiation								S					S
Solar wind plasma				s								s	
Ionospheric plasma				S			e	e					
Aurora plasma				E									
Trapped radiation	E	Ep	E			E							
Galactic cosmic ray		P	P										
Solar proton events	Sp	Sp	Sp										
Meteoroids							s				SEP	S	
Debris											s		
Dust											sep		

Upper-case letters are major effects and lower-case are minor effects. Assessment varies depending on spacecraft design

Table 3.3 Examples of different placement and orientation techniques for mitigating spacecraft/environment interactions

Placement	(1)	Place systems close to or far away from thermal sources for heat control
	(2)	Place radiation sensitive systems far from radiation sources such as an RPS, reactor, or RHUs
	(3)	Place contamination-sensitive devices out of the line of sight of known contamination sources
	(4)	Orient optical sensors flying in low Earth orbit so that they don't look over surfaces prone to glow
	(5)	Put radiation and meteoroid sensitive systems as close as possible to the center of the shielding protection system
Orientation	(1)	Orienting large, flat surfaces relative to the velocity vector in low altitude orbits to maximize or minimize drag.
	(2)	Orienting the more meteoroid or debris impact sensitive surfaces away from the maximum anticipated angle of fluence
	(3)	Orientation of current loops or the spin axis relative to the magnetic field to control magnetic torquing
	(4)	Orientation of oxygen erosion sensitive surfaces away from the vehicle normal while in low Earth orbit
	(5)	Orientation of a large array or space tether relative to the magnetic field and velocity vector to alter the induced electric fields
	(6)	Orientation so that sunlit and shadowed surface combinations that may cause arc discharges are minimized or that differential charging due to shadowing is minimized
	(7)	Orientation of a thermally sensitive surface in or out of sunlight to enhance heating or cooling

Table 3.4 Steps for selecting materials to limit environment interactions

Material selection	(1)	For charge mitigation, assess conductivity of internal and external materials
	(2)	To limit radiation effects on materials, assess the long term radiation response of the materials, particularly materials directly exposed to environment and those that are lightly shielded (e.g., behind thermal blankets)
	(3)	Assess the degradation from meteoroid, debris, and dust hypervelocity impacts
	(4)	For thermal control, assess absorptivity, emissivity, and transparency of materials
	(5)	Avoid materials with adverse outgassing properties
	(6)	Assess compatibility of materials at interfaces

Table 3.5 Difference between Class S and Class B parts [13]

Issue	Class S	Class B	Impact
Wafer lot acceptance	Required	–	Uniformity and pedigree traceability
Certification of production facilities	To specific assembly lines	To technologies and general facilities only	Burn-in and screening value relates to consistency of original product
Precap internal inspection	100 %	Sampled	Significant driver on level of reliability—criteria much more stringent in MIL-M-38510H
Particle impact noise detection (PIND) testing for loose particle detection	Required	–	Loose metallics in zero g field can cause failures
Serialization	Required	–	Traceability lost
Interim electrical test between test phases	Required	–	Potential of passing over problems and their causes
Burn-in	240 h	160 h	Later problem discovery
Reverse bias burn-in	Required	–	Impurity migration not detected
Interim electrical test after reverse bias burn-in	Required	–	Effects of reverse bias burn-in may be masked by subsequent actions
Radiographic inspection	Required	–	Observation of latent defects
Non-destructive 100 % bond pull test	100 %	Sampled	Parts with mechanical deficiencies get into equipment

their parts properties drift out of specification. Error detection and correction (EDAC) software is of special value in mitigating the effects of memory bit flips (SEU). Memory sparing and scrubbing are also useful in protecting against single event upset (SEU), single event latchup

(SEL), and generic part failures. Designing to worst-case parametric degradation values over voltage, temperature, life, and radiation is also of value. Where possible, a designer should introduce system redundancy at functional, subsystem, or system levels. Single-point failures and

effects need to be given special consideration in the system design. It is strongly recommended to perform a failure mode, effects, and criticality analysis (FMECA), parts stress analysis, worst-case analysis, and/or voltage, temperature, and frequency margin tests. In the case of spacecraft charging, the NASA Handbook 4002A [12] may be followed. Finally, all spacecraft circuitry should be analyzed and tested for electromagnetic compatibility (EMC) and electromagnetic interference (EMI), EMC/EMI.

3.4.6 Grounding

Proper grounding methods need to be considered in any electrical design. Standard techniques are described in NASA Handbook 4001 [14]. Typical methods are to provide a ground reference or resistive bleed path for circuit elements at all times—designers should choose system electrical and electronic grounding architecture to avoid structure currents and ground loops.

3.4.7 Trajectory

An obvious method for limiting environmental concerns is through careful orbital trajectory selection. A mission planner could consider optional trajectories that minimize meteoroid, radiation, and charging exposure. The Juno mission is a case in point because the spacecraft trajectory was selected to be highly eccentric and to pass over the poles in such a way that for much of the mission it avoids Jupiter's intense radiation belts. Similarly, the Voyager and Cassini trajectories were selected to pass through the gaps in the Saturnian rings to avoid particle impacts.

3.4.8 Operational Procedures

As in careful trajectory selection, operational procedures can significantly mitigate environmental effects. Specific procedures can be simulated using a testbed and the effects of part, assembly, functional, and spacecraft subsystem failures can be evaluated versus mission timeline. Fault tree analyses can be used to develop fault protection software. Vehicle operational procedures, such as orientation (say, relative to the Sun for thermal protection), can be implemented to limit specific effects during certain mission phases. For example, Cassini was flown in a fixed orientation during the cruise phase to provide a reduction in impacts on its rocket nozzles by a factor of approximately four. Finally, operational modes like hot (electronics 'on') versus cold (electronics 'off') and sparing can be used.

3.4.9 Construction Methods

Proper construction methods are a clear necessity for high reliability and for preventing design failures. These go hand in hand with correct handling procedures (see ISO 9000 practices for proper handling techniques). For complex space systems, the engineer must be acutely aware of how the electrical harnessing layout is constructed and the layout of grounding wires as these can contribute to ground loops and EMC/EMI concerns. Construction techniques need to minimize/limit the contamination of sensors, optics, paints, and coatings. A particularly dramatic example occurred for the first Shuttle tether experiment, on-board STS-75 in 1996. The electrical conductor of the tether was a copper braid wound around a nylon string, encased in Teflon-like insulation, with an outer cover of kevlar. All of this was then placed inside a nylon sheath. Apparently during construction, in winding up the tether on a spool, a wire filing was inadvertently forced into the cable insulation. As the tether was deployed, the imperfection caused a short between the Shuttle and the tether causing an arc discharge that severed the tether.

3.4.10 Interactions Versus Design Options

Given the various interactions and the design options available, a correlation matrix can be constructed that relates the appropriate effect with a means for mitigating that effect. Table 3.6 is such a matrix and provides an example of how the engineer might weight the comparative values of a design option versus specific interactions. Note that a '3' represents a principal method for mitigating an effect and should be given careful consideration in mission design, whilst '1' represents a method with minor effect. In an actual case, a mission-specific assessment should be made for the various spacecraft designs being evaluated.

3.5 Design Factors

The final step in determining the proper mix of design options and mitigating techniques for a given mission is to identify mission-specific design factors that must be considered by a project. While these factors, described in the following, are straightforward, the project management must carefully weigh their comparative value or impact on a specific project—they are not necessarily strictly engineering issues. As an example, the use of a radioisotope power source demands a strong technical justification, and the mission must undergo the Nuclear Safety Launch Approval process. A project must carefully weigh the advantages of the power source versus the additional costs of using it *in lieu* of a solar array.

Table 3.6 Design option space versus interactions, note assessment depends on spacecraft design

Interactions	Design options								
	Shielding	Positioning	Material properties	Electronic parts	Circuit/system design	Grounding	Trajectory	Operational procedures	Construction methods
Cumulative radiation effects	3	3	2	3	3		3	<i>1</i>	
Single event upsets	2	<i>1</i>		3	3		2	2	
Latch-up	2	<i>1</i>		3	3		2	2	
Surface charging/wakes	3	3	3	<i>1</i>	3	3	3	2	3
Internal charging	3	2	3	2	3	3	3	2	3
Power loss	3	3	3						<i>1</i>
$\mathbf{V} \times \mathbf{B}$		2				3	2		
Surface damage		3	3						<i>1</i>
Contamination	3	3	3			2		2	3
Glow		3					2	3	<i>1</i>
Particle impacts	3	3	2				3		
Torques		3	3				2	3	
Thermal	3	3	3				<i>1</i>	3	

A '3' represents a principal method for mitigation, whilst '1' represents a method with minor effect

3.5.1 Cost

The most obvious mission factor to be considered is cost in its various forms. In evaluating every design trade, there are inevitably monetary costs to be compared. The monetary costs are of course very dependent on the mission requirements and the class of the mission—typically, the higher the required reliability of the mission (Class A being the highest, Class D or now E being the lowest), the higher the monetary costs.

3.5.2 Mass

Most missions, because of launcher constraints, are driven by limited mass requirements. Radiation shielding (Europa) or the requirement for redundant systems (Pluto) would likely drive mass requirements in the environmental effects arena.

3.5.3 Power

Although power requirements and the power source (say, solar arrays versus radioisotope power sources) would be major issues for many programs, only the decision on

operating voltage will likely have a direct impact on environmental interactions. Many of the other design trades will be affected but will likely not be as critical for the space environment effects considered.

3.5.4 Complexity

Increased complexity may well be a major fallout from any environmental survivability trade. Key impacts would be in more elaborate shielding design (both radiation for Europa and thermal protection for Solar Probe), careful positioning (placement of systems on Europa and Solar Probe would be particularly constraining), special material selection (the Solar Probe heat shield), advanced EDAC software (the beacon mode reliability), enhanced redundancy (Pluto), and elaborate circuit design (Europa radiation hardening).

3.5.5 Reliability

Reliable operations over a decade or more (for example, the Pluto New Horizons mission) or in extreme environments (like the harsh radiation environment at Europa) require special care in the areas of EDAC software, redundancy,

and parts hardness. Long-term reliability ultimately leads to the need for a complete systems approach.

3.5.6 Availability

Availability encompasses multiple issues. The first is the well-known issue of parts or design availability—can the necessary parts or a usable design be found? The second issue is system availability to the operators—if this design fix is employed, will it lead to increased down time? That is, if the spacecraft is required to point a certain way to avoid meteoroid impacts would that limit the useful scientific data received?

3.5.7 Usability

In the case of usability, after applying a particular design trade, the designer should determine how it would influence the ease of operation of the vehicle. Would the design fix make it impossible to perform a certain series of operations? Would it rule out operations that might be critical to meeting the mission requirements? Furthermore, the designer should consider whether it is desirable to allow a spacecraft to do something potentially harmful to itself, or whether it is better to simply protect against this through diligent operations procedures. The use of operations procedures in place of physical limitations typically increases the likelihood of recovery from non-nominal scenarios, hence increasing usability.

3.5.8 Special Issues

Although a catch-all, the primary ‘special issues’ being considered here are the politics of radioisotope power supplies, their environmental issues, and the launch vehicle limitations imposed by such programs. Other examples are missions such as Europa and Pluto that place unique planetary contamination requirements on the spacecraft that may affect specific environmental design choices.

3.5.9 Design Options Versus Design Factors

Given the various design options available, a correlation matrix can be constructed that relates the design options and factors. Table 3.7 compares the overall trade space or set of design options with the various factors assuming a representative mission set (Europa, Pluto, and Solar Probe). The specific ratings provided are for a representative mission set where the major design factors would likely be radiation

and meteoroid shielding (Europa), material development (Solar Probe heat shield), software development (for a long-term autonomous beacon mode), redundancy (for the Pluto mission), parts hardness (Europa), and trajectory (all three missions would require complex trajectory calculations).

3.6 Designing for Survivability

To summarize, environmental interactions can have serious negative consequences for a mission’s survivability. Systematic consideration of the available mitigation techniques can limit these problems and lead to a much more reliable and often less expensive design. Tracing the pathway from environment to interaction to design options and then evaluating the options based on programmatic factors, however, can be an involved process. The rewards, though, should be obvious—a better-optimized design based on cost, mass, and reliability trades. At the least, by following such a process, the major space environment concerns can be identified early in a program when mitigation can be done relatively inexpensively. This chapter has provided a systematic method for considering the many trade-offs that need to be included. In particular, filling in Tables 3.2, 3.6, and 3.7 provides a formal means of carrying out a first-order evaluation of the ‘tallest tent poles’ in the optimization of a spacecraft design and provides a starting point for identifying the principal mitigation methods.

Table 3.1 summarizes the overall procedure for carrying out an integrated spacecraft design to optimize survivability in the space environment. The principal point to take away is that the designer must consider all the possible environments of concern and their effects early in the design process. Failure to do this can significantly increase the cost and schedule in developing a viable mission concept—it has been said that if finding and addressing an issue in the design phase costs one USD, addressing it in the construction phase will cost ten USD, while addressing it during launch preparations will cost 100 USD. Having to address it during flight may mean loss of the mission.

3.7 Suggested Resources for Space Environment and Survivability

To conclude this chapter, some of the main published resources that the reader should consider in developing a space environment survivability evaluation of a spacecraft are listed below. These are primarily reference books aimed at summarizing the environments and their effects in broad terms. WIKIPEDIA and other online sites are also listed. Unfortunately, and fortunately, these latter sites are periodically updated and thus are subject to change. Of particular

Table 3.7 Design options versus factors/criteria that must be considered in selecting between the options, note assessment depends on spacecraft design

Design options	Factors							
	Cost	Mass	Power	Complexity	Reliability	Availability	Useability	Special issues (RPS)
Shielding	3	3	2	3	3		3	3
Positioning	2	1	1	3	1	2	3	3
Material properties	3	1		2	2	3	1	3
Electronic parts	2			2	3	2	3	
Circuit/system design	3	2	3	3	3	2		3
Grounding	2		2	3	2		2	
Trajectory	1		2	2	2			3
Operational procedures	3		1	2	2		3	3
Construction Methods	1		2	3	2		1	2

A '3' represents a major, whilst '1' represents a minor effect

note, however, are the Space Environments and Effects homepage supported by NASA MSFC and the Space Environment Information System (SPENVIS) website supported by ESA. These two sites provide access to key environment and interaction programs for actually computing the environmental properties and their effects on space missions. Finally, the author strongly recommends that all environmental and survivability analyses start with a visit to your organization's Reliability Engineering professionals.

3.7.1 Further Reading

- Garrett, H.B., and C.P. Pike, eds. "Space Systems and Their Interactions with Earth's Space Environment." Prog. Astronaut. Aeronaut. 71, 1980.
- Jursa, A., ed. Handbook of Geophysics and the Space Environment. National Technical Information Services Document, Accession No. ADA 167000, 1985 [4].
- DeWitt, R.N., D.P. Dutson, and A.K. Hyder, eds. The Behavior of Systems in the Space Environment. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1994.
- Tribble, A. The Space Environment: Implications for Spacecraft Design. Princeton, NJ: Princeton University Press, 1995.
- Hastings, D., and H.B. Garrett. "Spacecraft-Environment Interactions." Atmospheric and Space Science Series, ed. A.J. Dessler. Cambridge, England: Cambridge University Press, 1996.
- Pisacane, V.L. "The Space Environment and its Effects on Space Systems", AIAA Press, Reston, VA, 2008.
- Garrett, H. B., and Whittlesey, A. C., "Guide to Mitigating Spacecraft Charging Effects", JPL Space Science and Technology Series, J. H. Yuen, Editor-in-Chief, John Wiley and Sons, Inc., Hoboken, NJ, 221 pages, 2011. (Note: also published as "Mitigating In-Space Charging Effects—A Guideline", NASA-HDBK-4002A, 3 March 2011.)

3.7.2 Further Online Reading

Note that these sites and addresses are subject to change; hence, the title could also be used as a search term.

Title (search term)	Address
Space Environment Information System (SPENVIS)	http://www.spennis.oma.be/
CRÈME Homepage (CREME96)	https://creme.isde.vanderbilt.edu/
NASA Technical Standards Program	http://standards.nasa.gov/
Space Engineering Standards (JPL)	http://engineer.jpl.nasa.gov/standards.html
Space Engineering Practices (JPL)	http://engineer.jpl.nasa.gov/practices/
Geomagnetic Field Models	http://www.ngdc.noaa.gov/geomag/
International Geomagnetic Reference Field	http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html
International Meteor Organization Index	http://www.imo.net/
Debris Models	http://orbitaldebris.jsc.nasa.gov/
Government—Industry Data Exchange Program (GIDEP)	http://www.gidep.org/
NASA National Space Science Data Center (NSSDC)	http://nssdc.gsfc.nasa.gov/
Today's Space Weather	http://www.swpc.noaa.gov/
The NASA Space Weather Bureau	http://spaceweather.com/
National Geophysical Data Center	http://www.ngdc.noaa.gov/
USGS Geomagnetism Program	http://geomag.usgs.gov/
The Aurora	http://www.geo.mtu.edu/weather/aurora/

(continued)

(continued)

Title (search term)	Address
DMSF Auroral Photos (Latest Aurora)	http://www.ngdc.noaa.gov/dmsp/
Recent Satellite Outages and Failures	http://sat-nd.com/#FAILURES
Meteor Showers	http://www.meteorblog.com/
NASA MSFC Space Environments and Effects (SEE)	http://see.msfc.nasa.gov/
JPL Homepage	http://www.jpl.nasa.gov/
ESA Space Debris	http://www.esa.int/esaMI/Space_Debris/

Acknowledgments The research and work that supported this chapter were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. Anon., *Guide to Reference and Standard Atmosphere Models*, AIAA G-003C-2010. Reston, VA: American Institute of Aeronautics and Astronautics, 2010.
2. Anon., *U. S. Standard Atmosphere, 1976*, NOAA-S/T 76-1562: NOAA, NASA, and USAF, 1976.
3. J. E. P. Connerney, "Magnetic fields of the outer planets," *J. Geophys. Res.*, vol. 98, pp. 18,659–18,679, 1993.
4. A. Jursa, "Handbook of Geophysics and the Space Environment." NTIS Document, Accession No. AD-A167000: AF Geophysics Laboratory, USAF, 1985.
5. Anon, *ASTM E490 - 00a(2006) Standard Solar Constant and Zero Air Mass Solar Spectral Irradiance*, American Society for Testing and Materials, 2000. DOI: [10.1520/E0490-00AR06](https://doi.org/10.1520/E0490-00AR06).
6. J. I. Minow, A. Diekmann, and W. Blackwell, Jr., "Status of the L2 and Lunar Charged Particle Environment Models," in *The 45th AIAA Aerospace Sciences Meeting and Exhibit*. AIAA paper 2007-0910, Reno, NV, 2007.
7. V. A. Davis, M. J. Mandell, D. L. Cooke, and D. C. Ferguson, "Nascap-2 k Spacecraft-Plasma Environment Interactions Modeling: New Capabilities and Verification," in *45th AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, 2007, p. 17.
8. J. Feynman, G. Spitale, J. Wang, and S. Gabriel, "Interplanetary Proton Fluence Model: JPL 1991," *J. Geophys. Res.*, vol. 98, pp. 13,281-13,294, 1993.
9. H. McNamara, R. Suggs, J. Jones, W. Cooke, and S. Smith, "Meteoroid Engineering Model (MEM): A Meteoroid Model for the Inner Solar System," *Earth, Moon, and Planets*, vol. 95, pp. 123.139, 2004.
10. N. Divine, "Five Populations of Interplanetary Meteoroids," *J. Geophys. Res.*, vol. 98, pp. 17,029-17,048, 1993.
11. J.-C. Liou, M. J. Matney, P. D. Anz-Meador, D. Kessler, M. Jansen, and J. R. Theall, "The New NASA Orbital Debris Engineering Model ORDEM2000," NASA/TP—2002-210780, May, 2002.
12. H. B. Garrett and A. C. Whittlesey, "Mitigating In-Space Charging Effects—A Guideline," NASA, Washington, DC NASA-HDBK-4002A, p. 181, Mar. 3, 2011.
13. Anon, *Class S Parts in High Reliability Applications*, NASA Preferred Reliability Practices, Practice No. PD-ED-1203.
14. A. C. Whittlesey, "Electrical Grounding Architecture for Unmanned Spacecraft," NASA, Washington, DC NASA-HDBK-4001, p. 29, Feb. 17, 1998.

Malcolm Macdonald

By definition astrodynamics is a truly modern field of engineering, the study of which dates back only as far as the early pioneers of space technology, such as Konstantin Tsiolkovsky (1857–1935). It was only realized as a practical field within engineering in the middle of the 20th century, as discussed in [Chap. 1](#), with the onset of the Space Age.

A concise definition of ‘astrodynamics’ is important to ensure sufficient distinction from related topics. Consider that ‘celestial mechanics’ is *a branch of astronomy concerned with the study of the motion of celestial objects*, ‘orbit dynamics’ is concerned with *the study of all orbiting bodies*, and ‘attitude dynamics’ is concerned with *the orientation of an object in space*. Meanwhile from [1], astrodynamics is defined as

the study of the motion of man-made objects in space subject to both natural and artificially induced forces.

Thus, astrodynamics combines features of each of the related ‘parent’ fields of science, and through the addition of “*artificially induced forces*” transposes these parent sciences into the field of engineering.

4.1 Introduction to Orbit Dynamics

An orbit is a geometric curve with no reference to time and can be either open or closed. A trajectory is the sequence of points in time along an orbit; the equations of motion of a body thus propagate the initial conditions along a trajectory.

M. Macdonald (✉)

Advanced Space Concepts Laboratory, Strathclyde Space Institute, University of Strathclyde, Glasgow, Scotland
e-mail: malcolm.macdonald.102@strath.ac.uk

The trajectory of a celestial body will diverge from its nominal orbit due to perturbations not considered when developing the nominal orbit.

4.1.1 Kepler’s Laws

Aristotle (384BC–322 BC) taught that circular motion was the only perfect motion. It followed therefore that this would be the motion of all ‘heavenly’ bodies. However, the work of a Danish nobleman, Tycho Brahe (1546–1601) and his assistant, the prematurely born son of a mercenary and a healer later tried for witchcraft, were to change this thinking forever. Brahe is credited with making the most accurate astronomical observations of his time and it is from these observations that his assistant would develop his own theories of ‘heavenly’ motion. In 1609, Brahe’s assistant published his first two laws of planetary motion. A third followed in 1619, and with this Johannes Kepler (1571–1630) had his three laws of planetary motion. Kepler’s laws are

1. The orbit of a planet is an ellipse with the Sun at a focus.
2. The rate of description of area by the radius vector joining planet to the Sun is constant.
3. The cubes of the semi-major axis of the orbit are proportional to the squares of the period of revolution.

Kepler’s first law describes the shape of an orbit, while also locating the central body. The second law tells how the angular velocity of a body changes with distance from the central body as it progresses around its orbit, and that the angular velocity is greatest at periapsis and least at apoapsis. Finally, the third law relates the size of an orbit to the period of revolution. From Brahe’s observations, Kepler’s laws are exact. Today they remain highly accurate approximations for the vast majority of natural celestial bodies and for spacecraft in weakly perturbed orbits. The term Keplerian motion is used to describe motion that exactly satisfies Kepler’s laws, but it should be noted that these laws are only a description of the motion, not an explanation.

4.1.2 Bode's Law and Commensurabilities

A curious feature of orbit dynamics is the large number of apparently coincidental relationships that can be defined between bodies within the same system. One particular relationship of note is Bode's Law (also called Titius–Bode law), first published in 1772. This gives the mean distance from the Sun of each of the planets in our solar system as

$$r_n = 0.4 + 0.3(2^n) \quad (4.1)$$

where n takes the values $-\infty, 0, 1, 2, 3$, and so forth. When published, Bode's law approximately satisfied all of the known planets, with a gap between the fourth planet (Mars) and the fifth planet (Jupiter). Furthermore, when Uranus was discovered in 1781 it was found to also satisfy Bode's law. Consequently, attention was drawn to the apparent gap between the fourth and fifth planets, and subsequently Ceres, the largest object in the asteroid belt, was found at Bode's predicted distance in 1801. Bode's law was widely accepted until the discovery of Neptune in 1846, which was found not to conform to the law.

The exact explanation of the apparent accuracy of Bode's law is unclear. Similar to Bode's law, a notable number of commensurabilities exist, both in our solar system and in other orbital systems. For example, a 1:2:4 resonance in orbital period is exhibited by Jupiter's moons Ganymede, Europa and Io. A second example of such an orbital resonance, that is a 1:2:4 ratio, was confirmed in 2010 in the extrasolar planets Gliese 876c, Gliese 876b and Gliese 876d [2]. Once again, the exact explanation for such commensurabilities remains unclear; a detailed discussion of some commensurabilities can be found in [3].

4.1.3 The Two-Body Problem

Given at any time the positions and velocities of two massive particles that are moving under their mutual gravitational force alone, the mass of each being known, the two-body problem seeks to determine their positions and velocities at any other time. This problem was first posed and solved by Isaac Newton (1642–1727). The two-body problem is important because it is the only gravitational problem in dynamics for which a complete and unconstrained general solution can be defined. Furthermore and as with Kepler's laws, a wide range of practical orbital problems can be approximated as two-body problems, providing approximate solutions to these problems.

4.1.3.1 Newton's Laws of Motion

Providing the explanation and physical rationale for Kepler's laws, Newton introduced his three laws of motion in his

1687 manuscript, *Philosophiæ Naturalis Principia Mathematica*, or simply *Principia*. These are

1. Every body continues in its state of rest or uniform motion in a straight line except in so far as that state is compelled to change by forces impressed on it.
2. The rate of change of momentum of a body is proportional to the impressed force and takes place in the direction in which that force is impressed.
3. To every action there is always an opposed equal reaction: or, mutual actions of two bodies upon each other are always equal and directed in contrary parts.

4.1.3.2 Newton's Law of Universal Gravitation

Together with Newton's second law of motion, when applied to a constant-mass system, Newton's law of universal gravitation provides the basis for celestial mechanics and astrodynamics. As with Newton's laws of motion, Newton's law of universal gravitation was introduced in *Principia* and can be stated as: Every particle of matter in the universe attracts every other particle of matter with a force directly proportional to the product of the masses and inversely proportional to the square of the distance between them. Hence, for two particles of mass m_1 and m_2 separated by a distance r , the mutual force of attraction, F , is

$$F = G \frac{m_1 m_2}{r^2} \quad (4.2)$$

where G is the gravitational constant, often called the universal gravitation constant. It should be noted that the universal gravitation constant is widely regarded as the most difficult physical constant to accurately measure; the 2010 Committee on Data for Science and Technology (CODATA) recommended value of the gravitational constant has a relative standard uncertainty of 1.2×10^{-4} .

Some problems do exist with Newton's law of universal gravitation. For example, as Newton himself noted the law requires the force to act instantaneously, in a vacuum, and without the mediation of anything by or through which the force could act; commonly referred to as 'action at a distance'. In many ways, Newton's law of universal gravitation has been superseded by Albert Einstein's (1879–1955) theory of general relativity, which attributes gravity to curved spacetime instead of a force propagated between two bodies. In general relativity masses distort nearby spacetime, with other particles thereby moving in trajectories determined by the geometry of spacetime. However, Newton's law of universal gravitation provides an excellent approximation for the effect of gravity in non-relativistic situations and therefore remains of great value.

4.1.3.3 Solution of the Two-Body Problem

Assuming each of the two bodies are point masses and that each feels no force other than the mutual gravitational

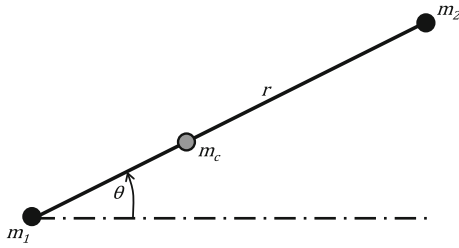


Fig. 4.1 Physical relationship between masses, with indicative center of mass shown

attraction, the mutual force can be equated using Newton's third law. Assuming an inertial reference frame and using Newton's second and universal gravitation laws it can be shown that the center of mass of the two-body system is at rest, or moves with constant velocity, forming a new inertial frame. Thereafter, the equations of motion of each mass can be written as a single equation, the two-body equation of motion

$$\ddot{\mathbf{r}} + \frac{\mu}{r^3} \mathbf{r} = 0 \quad (4.3)$$

where $\mu = G(m_1 + m_2)$ is called the gravitational parameter and \mathbf{r} is the vector joining m_1 and m_2 , as shown in Fig. 4.1. Note that if $m_1 \gg m_2$ then the gravitational parameter is typically written as $\mu = Gm_1$; this is especially convenient as μ can be determined to a high degree of accuracy through trajectory observation, thereby negating the low level of accuracy in knowledge of the universal gravitation constant.

From Eq. 4.3 it can be shown that the total energy is conserved for the two-body problem, that the motion is in a plane normal to the angular momentum vector, and that the angular momentum of the system is conserved, with the angular momentum vector being twice the rate of description of area by the radius vector. This final point is Kepler's second law in a mathematical form.

The two-body equation of motion, Eq. 4.3, can be solved to obtain the position of a particle as a function of its position around the orbit, where position is used as the independent variable in place of time. Solving for position yields

$$r(\theta) = \frac{h^2/\mu}{1 + (Ah^2/\mu) \cos \theta} \quad (4.4)$$

where A is a constant of integration and h is the orbit angular momentum. It should be apparent that Eq. 4.4 is the polar equation of a conic section and thus that all orbits are conic sections and may be better written as

$$r = \frac{p}{1 + e \cos \theta} \quad (4.5)$$

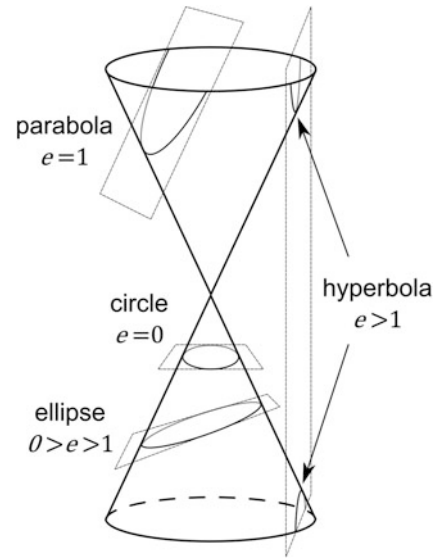


Fig. 4.2 Types of conic section with associated orbit eccentricity range. Image Malcolm Macdonald

where $p = h^2/\mu = a(1 - e^2)$ is the orbit semi-latus rectum, $e = Ah^2/\mu$ is the orbit eccentricity, and θ is the orbit true anomaly, also often written as v in celestial mechanics literature; each of which will be discussed later. The solution of the two-body problem, a conic section, provides a mathematical rationale for Kepler's first law as a special case of the solution in which the orbit eccentricity is less than one.

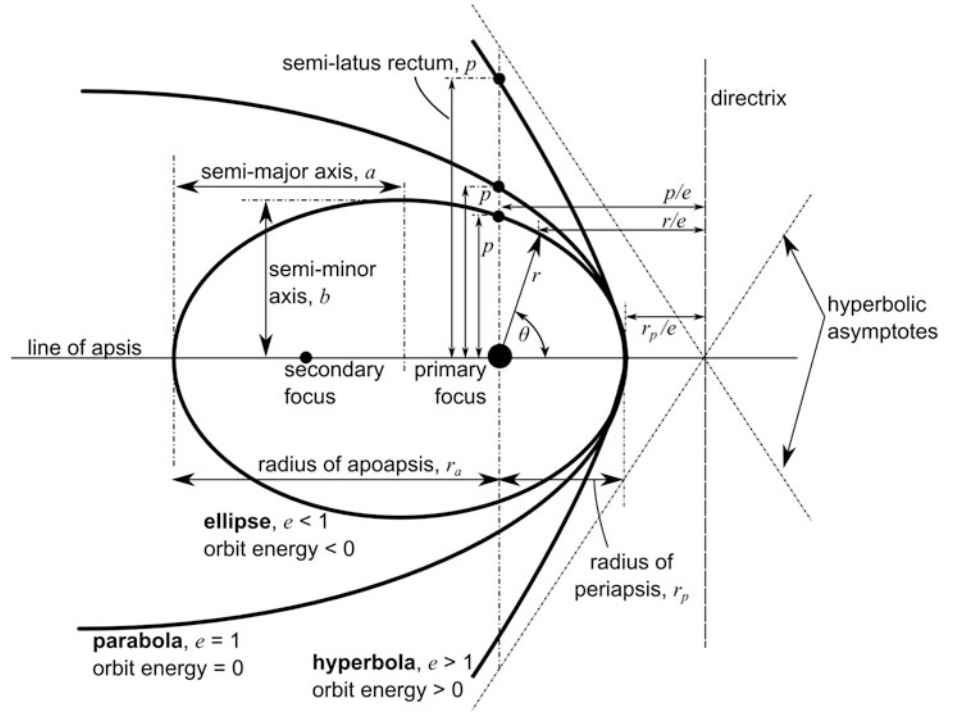
4.1.3.4 Conic Sections

Conic sections are the curves generated by the intersection of a plane with one or two nappes of a cone. All orbits are conic sections as shown in Fig. 4.2. A plane that fully intersects only a single nappe produces an ellipse, eccentricity less than one; if the plane is perpendicular to the axis of the cone, it produces a circle, eccentricity equal to zero; a special case of the ellipse. Such orbits are closed periodic returning orbits. If the plane does not fully intersect a single nappe it produces a parabola, eccentricity equal to one, and if it intersects both nappes it produces a hyperbola, eccentricity greater than one. Such orbits are open, non-returning (or escape) orbits.

4.1.4 Basic Orbit Parameters

All conic sections have two focal points, or foci. In astrodynamics the gravitational center of attraction sits at one focus, termed the primary focus. In the special case of a circular orbit the two foci are coincidental, while in a parabola orbit one focus is removed to infinity. As shown in Fig. 4.2, the hyperbola has a branch associated with each

Fig. 4.3 Conic section parameters. *Image* Malcolm Macdonald



focus. The case of two separate foci, as in an elliptical orbit is shown in Fig. 4.3.

The major and minor axes determine the size and shape of a conic section. Typically, half sizes are used and termed as such. Hence as shown in Fig. 4.3, the orbit size and shape is regulated by the semi-major axis, a , and semi-minor axis, b , where these are equal in the special case of a circular orbit. The extreme points of an ellipse are the periapsis, or pericentre, and apoapsis, or apocentre, the closest and furthest points of an orbit respectively. The terms periapsis and apoapsis can be altered to indicate the central body at the primary focus, for example in Earth orbit the terms become perigee and apogee, while in a solar orbit they become perihelion and aphelion. The location on an orbit is defined as the true anomaly, θ , also often written as ν in celestial mechanics literature, and is measured as the angular displacement from the periapsis to the radius vector. The true anomaly varies from 0° to 360° and because it is measured from the periapsis is not defined for a circular orbit, where instead the true or mean longitude can be used. From Eq. 4.5, the radius of periapsis is found to be

$$r_p = a(1 - e) \quad (4.6)$$

whilst the radius of apoapsis is

$$r_a = a(1 + e). \quad (4.7)$$

The shape of an orbit can be characterized by the single parameter eccentricity, e , which by definition is always positive and is defined as

$$e = \frac{\sqrt{a^2 - b^2}}{a}. \quad (4.8)$$

Combining Eqs. 4.6 and 4.7 yields the further useful relationships

$$a = \frac{r_a + r_p}{2} \quad (4.9)$$

and

$$e = \frac{r_a - r_p}{r_a + r_p}. \quad (4.10)$$

4.1.4.1 Orbit Period

From Eq. 4.3 recall that, in the mathematical form, Kepler's second law may be given as stating that the angular momentum vector is twice the rate of description of area by the radius vector, that is

$$h = r^2 \dot{\theta}. \quad (4.11)$$

From this, Kepler's third law can be given in the mathematical form by noting that the area of the ellipse to be circumnavigated in time T is πab . Thus

$$h = \frac{2\pi a^2 \sqrt{1 - e^2}}{T} \quad (4.12)$$

and from Eq. 4.5

$$h^2 = \mu a(1 - e^2) \quad (4.13)$$

giving

$$T = 2\pi\sqrt{\frac{a^3}{\mu}}. \quad (4.14)$$

This shows that the orbit period depends on only the orbit size, that is the semi-major axis, and the sum of the masses in the system.

4.1.4.2 Orbit Velocity

The velocity V of a body at radius r acts tangential to the orbit and will hence have component \dot{r} along the radius vector and $r\dot{\theta}$ perpendicular to the radius vector. Thus

$$V^2 = \dot{r}^2 + r^2\dot{\theta}^2. \quad (4.15)$$

From Eqs. 4.5 and 4.11

$$\begin{aligned} V^2 &= \left(\frac{h}{p}\right)^2 [2 + 2e \cos \theta - (1 - e^2)] \\ &= \frac{2h^2}{rp} - \left(\frac{h}{p}\right)^2 (1 - e^2). \end{aligned} \quad (4.16)$$

Recalling that $p = h^2/\mu = a(1 - e^2)$, the traditional form of the *vis-viva equation*, also referred to as the orbital energy conservation equation, can be written as

$$V^2 = \mu \left(\frac{2}{r} - \frac{1}{a} \right). \quad (4.17)$$

It is a simple matter to reduce Eq. 4.17 to find the velocity on a circular orbit, where $r = a$, or to find the escape velocity, where $a \rightarrow \infty$, for a minimum energy escape trajectory.

4.1.4.3 Flight-Path Angle

The flight-path angle can be helpful in determining an effective cross-sectional area for use in perturbations analysis. The flight-path angle is measured from the local horizontal, defined as perpendicular to the radius vector, to the velocity vector. As such, the flight-path angle is zero when the orbit eccentricity is zero.

4.1.4.4 Orbit Energy

The energy of an orbit is the sum of the kinetic and potential energies of the orbiting bodies. If the two-body center of mass is not accelerating and assuming that $m_1 \gg m_2$, then the orbital energy of the larger body can be neglected and the orbital energy, E_T , can be defined in terms of the smaller body alone as

$$E_T = T + U = \frac{V^2}{2} - \frac{\mu}{r} \quad (4.18)$$

where, T is the kinetic energy per unit mass. Using Eq. 4.17 the orbit energy thus reduces to

$$E_T = -\frac{\mu}{2a}. \quad (4.19)$$

Note that the orbit energy is dependent on only the orbit semi-major axis and that only a hyperbolic orbit has positive orbit energy.

4.1.4.5 Semi-Latus Rectum

As the eccentricity tends towards unity, the orbit energy increases towards zero and the semi-major axis tends towards infinity. At eccentricity equal to one, that is, on a parabolic orbit, the orbit energy is zero and the semi-major axis tends to infinity. For this reason, it is often convenient to use the semi-latus rectum, p , parameter to describe an orbit because it remains defined for all eccentricity as illustrated in Fig. 4.3. The semi-latus rectum is the distance from the primary focus to the orbit, measured perpendicular to the line of apsis

$$p = \frac{h^2}{\mu} = \frac{b^2}{a} = a(1 - e^2). \quad (4.20)$$

Note that the semi-latus rectum is zero for all rectilinear orbits.

4.1.4.6 Rectilinear Orbit

A rectilinear orbit is a limiting case of all orbits, where the eccentricity is equal to one and occurs when the plane intersecting the cone is coincidental with the surface of the cone. In a rectilinear orbit, the radius of periapsis is zero such that, for example, a rectilinear ellipse becomes a line segment connecting both foci, while a rectilinear parabola and a rectilinear hyperbola are each a line from the focus, along the line apsis to infinity. In the rectilinear ellipse, the line is traversed with maximum velocity at one focus and zero velocity at the other, while in a rectilinear parabola maximum velocity occurs at the focus, with velocity tending to zero as the radius tends to infinity. In a rectilinear hyperbola some velocity remains as the radius tends to infinity.

4.1.4.7 Kepler's Equation

The radius vector sweeps through 360° , 2π radians, in one orbit period. The mean motion, n , or the mean angular velocity, is thus

$$n = \frac{2\pi}{T}. \quad (4.21)$$

Thereafter, defining the time t of periapsis passage as τ , the angle swept by the radius vector in time $(t - \tau)$ is defined as the mean anomaly, M

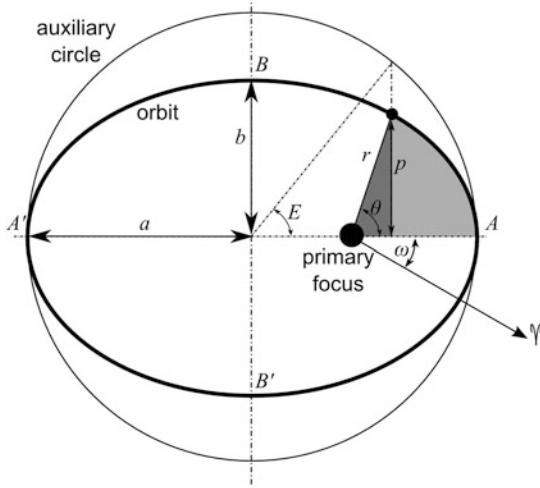


Fig. 4.4 Elliptical orbit parameters. *Image Malcolm Macdonald*

$$M = n(t - \tau). \quad (4.22)$$

Using an auxiliary circle on an elliptical orbit allows the eccentric anomaly, E , to be defined as shown in Fig. 4.4. The eccentric anomaly is related to the true anomaly by

$$\tan\left(\frac{\theta}{2}\right) = \left(\frac{1+e}{1-e}\right)^{1/2} \tan\left(\frac{E}{2}\right). \quad (4.23)$$

A complete derivation of this relationship can be found in [3].

The eccentric anomaly and mean anomaly are related by Kepler's equation, which ultimately relates the time and angular displacement around an orbit. Using Kepler's second law, the auxiliary circle and subdividing the area swept by the radius vector since periapsis passage into the two gray regions shown, Kepler's equation can be derived as

$$E - e \sin E = M = n(t - \tau). \quad (4.24)$$

A complete derivation of this relationship can be found in [3, 4] along with the semi-analytical solution to the equation.

4.1.4.8 Satellite State

Thus far, consideration of an orbit has been constrained to within the orbit plane. However, since the orbit exists within three-dimensional space, the location of the orbit plane must also therefore be defined; this complete set of information is termed the 'state' of the satellite in space and requires six quantities to be fully defined. The 'state' of the satellite can be defined in many ways, using different but equivalent forms that are called either a 'state vector', when comprised solely of scalar magnitude terms, usually three position and three velocity components, or an 'element set', when comprised of a set of geometric parameters, usually a mix of scalar magnitudes and angular representations. A

'state vector' or an 'element set' is always associated with a reference epoch, with time becoming, in effect, a seventh element of the set and always referring to a particular reference frame. It should be noted that time can also be used within the element set as the time of periapsis passage, but this does not negate the requirement for a reference epoch to give a position around the orbit.

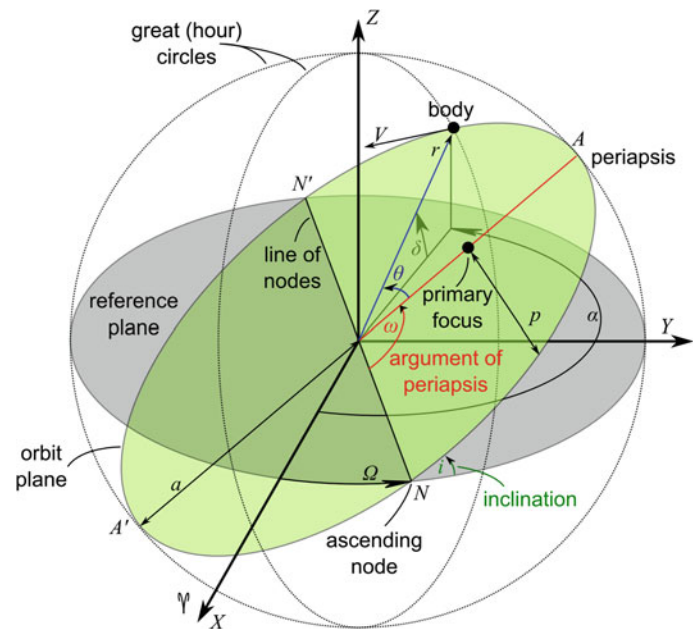
Element sets can take a number of forms due to the variety of orbital elements that can be used; the application domain typically defines which set is most appropriate. The most commonly used element set is called Classical Orbital Elements, also known as Keplerian Orbital Elements, and is illustrated in Fig. 4.5. The Classical Orbital Elements are defined within an inertial rectangular frame of reference centered at the primary focus with the X -axis directed towards the zero point of longitude, also often called the First Point in Aries even though, due to orbital precession, the vernal equinox is no longer within the constellation Aries. The zero point of longitude is an arbitrary fiducial direction in the reference plane at which right ascension, defined later in this section, is zero. The zero point of longitude is often defined as the northern vernal equinox in 1950 or 2000. However, as will be seen in Sect. 4.1.5 other fiducial directions can be used. It should be carefully noted which fiducial direction is used to define the zero point of longitude. The Z -axis is typically aligned positively along the spin axis of the central body towards the north pole. The frame of reference is completed in the right-hand sense by the Y -axis.

Despite several problems with their use in orbit propagation, which will be discussed later, the Classical Orbital Elements are very useful because they provide a direct representation of the shape, size and orientation of an orbit, as illustrated in Fig. 4.5. The Classical Orbital Elements allow the orbit plane to be located in space through three angular parameters

- Ω , the longitude, or right ascension of the ascending node measured from the X -axis to the ascending node, N , of the orbit within the reference frame and in a right-handed sense.
- ω , the argument of periapsis measured from the ascending node, N , of the orbit to the periapsis of the orbit within the orbit plane and in a right-handed sense.
- i , the inclination between the reference and orbit planes, measured from the reference plane in a right-handed sense. Note that inclinations $<90^\circ$ are called prograde; inclinations $>90^\circ$ are called retrograde, and orbits with inclination $=90^\circ$ are called polar orbits.

Although the semi-major axis, eccentricity, right ascension of the ascending node, argument of periapsis and inclination serve to define the orbit, a sixth element is required to define a position on the orbit at a given epoch. This sixth, position fixing element can be, amongst other things, the true anomaly or the mean anomaly.

Fig. 4.5 Keplerian orbit parameters. *Image Malcolm Macdonald*



It was noted earlier that the semi-major axis is ill-defined in certain circumstances, prompting the use of the semi-latus rectum, and that the true anomaly is not defined for a circular orbit. It should be further noted that the argument of periastron is also not defined for a circular orbit. Hence, the argument of latitude, $u \equiv \omega + \theta$, can be used in place of both the true anomaly and the argument of periastron and is measured from the ascending node to the position vector. Similarly, the right ascension of the ascending node is not defined for a zero inclination orbit. Hence, the true longitude, L , can be used to define the angle between the X -axis and the position vector. The classical orbital elements are summarized in Table 4.1.

4.1.5 Coordinate Systems

The use of an inertial rectangular frame of reference was discussed in the previous section. Such a reference frame at Earth is termed an Earth Centered Inertial, ECI, or Geocentric Inertial, GCI, reference frame with the Z -axis meeting the celestial sphere at the north celestial pole. This reference frame is also termed a Geocentric Equatorial Coordinate System, IJK. Note that at Earth the Z -axis, and hence the reference frame shown in Fig. 4.5, is angularly displaced from the plane of the ecliptic, the plane of the Earth's orbit around the Sun, through an angle known as the obliquity of the ecliptic, ϵ , equal to approximately 23.4° . An equivalent reference frame centered at the Sun is termed a Heliocentric Inertial, HCI, reference frame, also called a Heliocentric Coordinate System, XYZ. The correct choice of reference frame can often significantly reduce the

complexity of a problem; consider the motion of the planets in an Earth-centered and Sun-centered inertial frames.

In general, three types of coordinate system can be identified in astrodynamics by consideration of the location of the origin of the system. Specifically, an object's center of mass, as in Fig. 4.5, the system's center of mass, or barycenter, and a non-inertial rotating system using the barycenter, called a synodic system.

4.1.5.1 Position on Earth

Knowledge of an observer's location on Earth is critical to many problems in orbit dynamics, including accurately locating a ground station, or for remote sensing of Earth. Two coordinates, latitude and longitude define a location on the surface of a planet. Longitude is an angular displacement measured from a prime meridian. In the case of the Earth, it was proposed at the International Meridian Conference, held in Washington, D.C., in the United States of America, in 1884 that the prime meridian be, "the meridian passing through the center of the transit instrument at the Observatory of Greenwich as the initial meridian for longitude" [5]. This resolution passed 22–1, with San Domingo (now the Dominican Republic) voting against, and France and Brazil abstaining; the French did not adopt the Greenwich meridian until 1911. It was also proposed at the International Meridian Conference that "longitude shall be counted in two directions up to 180° , east longitude being plus and west longitude minus" [5]. This resolution was, however, rather controversial and provoked much discussion over the use of two directions rather than a single range from 0° to 360° . The Earth's axis of rotation is termed the 'poles' and the equator is the locus of points on the surface

Table 4.1 Definition of Classical Orbital Elements

Parameter	Symbol	Definition
Semi-major axis	a	Half the major axis of an orbit's ellipse
Semi-minor axis	b	Half the minor axis of an orbit's ellipse
Eccentric anomaly	E	$\cos^{-1}(\frac{1}{e}(1 - \frac{r}{a}))$
Eccentricity	e	$\frac{\sqrt{a^2 - b^2}}{a}$
Inclination	i	Angle between the orbital plane and a reference plane
Eccentric longitude	K	$\varpi + E$
True longitude	L	$\Omega + \omega + \theta = \varpi + \theta$, a broken angle, measured in the reference plane from the zero point to the ascending node and then around the orbit to the satellite
Mean longitude	l	$\varpi + M$, a broken angle, measured in the reference plane from the zero point to the ascending node and then around the orbit. <i>Nota bene</i> , the convention established in [3] is followed in this handbook, however in some literature mean longitude is denoted L , while true longitude is denoted l
Mean anomaly	M	$n(t - \tau)$
Mean motion	n	$\sqrt{\mu/a^3}$, the mean motion, or mean angular velocity
Semi-latus rectum	p	$a(1 - e^2)$, half a chord through the focus and parallel to the conic section directrix
Orbit radius	r	Distance from the coordinate system origin, typically coincident with the center of the central body, to the satellite
Argument of latitude	u	$\omega + \theta$, angle from the ascending node to the position vector
True anomaly	θ	$\cos^{-1}(\frac{1}{e}(\frac{r}{a} - 1))$, angle from periapsis to the satellite, measured within the orbit plane
Argument of periapsis	ω	Angle from the ascending node to the satellite when at periapsis, measured within the orbit plane
Longitude of ascending node	Ω	Angle between line of nodes and the zero point of longitude in the reference plane
Longitude of periapsis	ϖ	$\Omega + \omega$, a broken angle, measured in the reference plane from the zero point to the ascending node and then around the orbit to periapsis

created by the perpendicular plane to the axis of rotation passing through the center of mass. The Earth's equatorial plane is the reference plane shown in Fig. 4.5 and extends out from the equator; it is also the reference plane for measuring latitude, measured north–south from the reference plane with values from 0° to $\pm 90^\circ$; positive is the northern hemisphere.

Whilst a perfectly spherical Earth is often used to locate an observer on the surface, it is noted that the Earth is not a perfect sphere. However, such an assumption is sufficiently accurate for many initial studies. If a more accurate location is required, several models exist. In particular, for Earth a simple ellipsoid model works well; specifically an oblate spheroid with the semi-major axis equal to the equatorial radius and semi-minor axis equal to the polar radius. However, note that in the case of the Moon a triaxial ellipsoid representation works better. It should be apparent that longitude is by definition, compliant with a non-spherical Earth, but this is not the case for latitude.

The reference ellipsoid provides an approximation to the hypothetical surface denoted as the mean sea level. The actual mean sea level surface (if the oceans were in

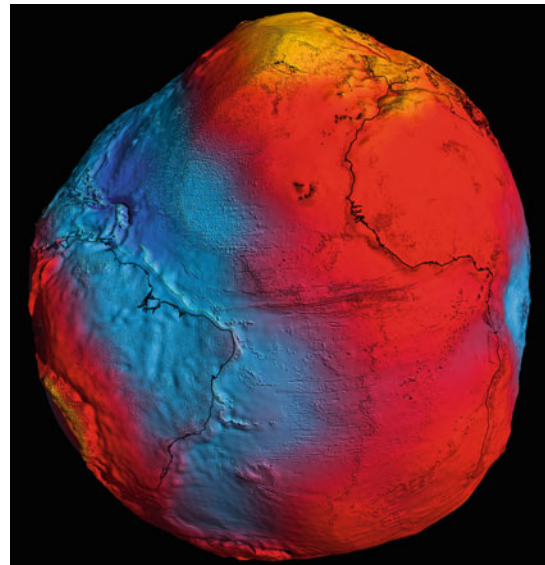
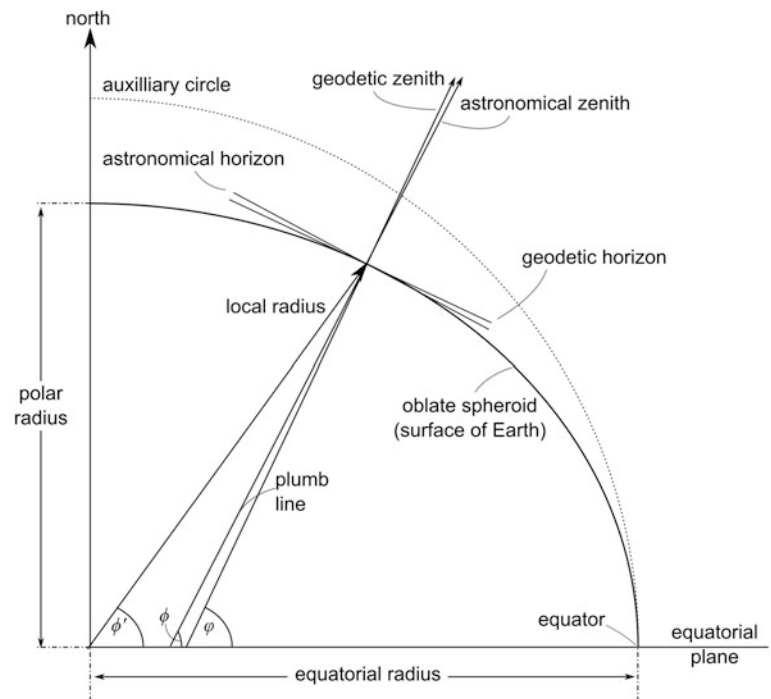


Fig. 4.6 Earth's geoid mapped from data collected by the gravity field and steady-state ocean circulation explorer (GOCE). Colors represent deviations in height (± 100 m) from an ideal geoid; blue colors represent low values and reds/yellows represent high values. Image ESA

Fig. 4.7 Latitude geometry, not to scale. *Image Malcolm Macdonald*



equilibrium, at rest relative to the rotating Earth) extended through the continents is called the geoid. The geoid deviates from the reference ellipsoid due to the uneven distribution of mass within the Earth, as illustrated in Fig. 4.6. A plumb line taken from an observer's location on the surface of the Earth will in general not intersect the Earth's equatorial plane at the center of the Earth due to the variation of the geoid from the reference ellipsoid. The plumb line will intersect the equatorial plane at an angle ϕ , termed the astronomical latitude. A line joining the observer to the center of the Earth intersects the equatorial plane at an angle ϕ' , termed the geocentric latitude.

The non-spherical Earth introduces a third definition of latitude, termed the geodetic latitude, ϕ , which is the intersection angle of a line normal to the surface of the ellipsoid with the equatorial plane. The latitude used on most maps is the geodetic latitude. The difference between the geodetic latitude and the astronomical latitude is typically very small, but is non-negligible for high-accuracy observations and is termed the 'station error' or 'deflection of the vertical'. The geometry of these latitudes is illustrated in Fig. 4.7.

4.1.5.2 Celestial Sphere

The celestial sphere is an imaginary sphere of arbitrary radius, which is concentric with the Earth and rotating upon the same axis. It can often be convenient to suppose that observed objects are simply located at fixed positions on the inside of this sphere, especially within astronomy when the distance to an object is of no concern. Alternatively, it can be supposed that objects move on the inner surface of the

sphere. The celestial poles are located at the intersection of the Earth's rotational axis with the celestial sphere. Great circles are the intersection of the celestial sphere by a plane passing through the center of the sphere, while hour circles are great circles that include the axis of rotation, as shown in Fig. 4.5.

The location of an object on the celestial sphere is described by two angular quantities. Similar to a position on Earth, this can be either the ecliptic (or celestial) latitude and longitude, which use the ecliptic as a reference plane and take the arbitrary fiducial direction as the First Point in Aries, or alternatively right ascension and declination, which use the equatorial plane and take the arbitrary fiducial direction as the vernal equinox at a given epoch. As illustrated in Fig. 4.5, right ascension, α , is measured positively from the fiducial direction to the east and within the reference plane, hence unlike longitude varies from 0° to 360° . Declination, δ , is measured from the reference plane positively to the north and negative to the south and varies from 0° to 90° . Hence, at Earth and on the celestial sphere the declination and terrestrial latitude are the same, but right ascension and terrestrial longitude differ because they use different fiducial directions.

4.1.5.3 Supplementary Coordinate Systems

In addition to the Earth Centered Inertial (Geocentric Inertial/Geocentric Equatorial Coordinate System) and the Heliocentric Inertial (Heliocentric Coordinate System), several other coordinate systems can be defined, such as the International Celestial Reference System, ICRS, a standardized reference system adopted by the International

Astronomical Union (IAU) with origin at the solar system barycenter. The primary fiducial direction of the ICRS frame is the IAU-76/FK5 (Fifth Catalog of Fundamental Stars) value adopted for the quasar 3C 273,¹ at the epoch J2000.0 (defined in Sect. 4.1.6). Other useful coordinate systems include the geocentric equivalent of the ICRS, the Geocentric Celestial Coordinate System, GCRF and the Body-Fixed Coordinate System, ITRF. Noting that a reference frame with origin at the center of the Earth is called geocentric, while a reference frame with origin on the Earth's surface is called topocentric, several other coordinate systems can be usefully defined. Topocentric systems will typically make use of azimuth and elevation angles to describe the location of a body. Azimuth is the angle measured from the north, in a clockwise sense, to a point immediately below the object of interest. Elevation is the angle measured from the local horizon upwards to the body of interest; as such, an object is only visible if it has an elevation of greater than zero. See [3, 4] for a detailed description of additional reference systems and conversions between systems and in terms of the celestial sphere.

4.1.5.4 Satellite-Based Coordinate Systems

A great number of satellite-based coordinate systems exist in the literature, however little is standardized and often a system is developed to fill a requirement of a specific mission.

Perhaps the most used satellite-based coordinate systems is the so-called Gaussian coordinate system or simply 'RTN' for radial, tangential and normal. The system can also be referred to as a local vertical, local horizontal, LVLH, system and is illustrated in Fig. 4.8 as axis *RSW*. The origin is located at the satellite with the radial vector always pointing away from the central body along the radius vector towards the satellite; note that by this definition the positive radial vector is zenith pointing and the negative radial vector is nadir pointing. The *W*-axis is directed along the cross, or vector product of the radius and velocity vectors. The *S*-axis completes the system in a right-handed sense. As such, the *S*-axis is only coincident with the velocity vector for circular orbits, or at crossings of the line of Coordinate Systems.

An alternative satellite-based coordinate system is shown in Fig. 4.8 as axis *NTW*, here the *T*-axis is always tangential to the orbit and hence directed along the velocity vector. Once again, the *W*-axis is directed along the cross, or vector product of the radius and velocity vectors. The *N*-axis completes the system in a right-handed sense. This system

is useful to analyze changes in orbit velocity due to, for example, atmospheric drag.

The third satellite-based coordinate system shown in Fig. 4.8 is axis *PQW*. It is actually a rotated geocentric system, and is convenient for the processing of remote sensing data. The *P*-axis is directed towards the periapsis, the *Q*-axis is perpendicular to the *P*-axis within the orbit plane and in the direction of orbit rotation. The *W*-axis completes the system in a right-handed sense. Further satellite-based coordinate systems can be found in [4].

4.1.5.5 Ground Track

The ground track, or trace, of a spacecraft is the locus of points generated by the spacecraft position vector as it intersects the reference ellipsoid that provides an approximation to the hypothetical surface denoted as the mean sea level. The ground track of three different orbits is shown in Fig. 4.9; each of which will be discussed in Sect. 4.4. Note that although Fig. 4.9 shows only closed orbits, that is ellipses, a ground track can be derived for all spacecraft. Furthermore, in Fig. 4.9 the three orbits shown complete an integer number of revolutions in a sidereal day (defined in Sect. 4.1.6), causing the ground track to revisit the same path over the surface. In general, it is preferable for an Earth orbiting spacecraft to complete an integer number of revolutions in an integer number of sidereal days, as this aids ground management and operations of the spacecraft. For example, European remote sensing spacecraft, including ERS-1, ERS-2 and ENVISAT, are typically inserted into orbits that repeat their ground tracks over a period of 35 sidereal days, completing 501 revolutions in that time. If the spacecraft does not complete an integer number of revolutions in an integer number of sidereal days then the ground track will never repeat.

Information on the size, shape and inclination of an orbit can be inferred directly from a ground track. For example, note that the ground track of each orbit in Fig. 4.9 does not exceed a latitude magnitude equal to the orbit inclination. Additionally, an orbit with a period of less than one sidereal day and of prograde inclination, i.e. $<90^\circ$, will predominantly move from west to east along its ground track. Because the ground track will move in the same direction as the Earth's rotation, this is termed 'apparent direct', or 'apparent prograde' motion. In contrast an orbit with a period greater than one sidereal day and of prograde inclination will predominantly move from east to west along its ground track. This is termed 'apparent retrograde' motion and results from the orbit moving over the surface of the Earth at an angular rate less than the rate of rotation of the Earth. The ground track of a retrograde orbit will always move predominantly from east to west, irrespective of the orbit period. Note further that due to variations in orbit eccentricity and/or inclination, it is possible for the ground

¹ 3C 273 was one of the first quasars discovered in the early 1960s, alongside 3C 48, and the first object to be identified as a quasi-stellar radio source, or "quasar", a very energetic and distant active galactic nucleus. 3C 273 is the optically brightest quasar in our sky (apparent magnitude, $m \sim 12.9$), and one of the closest with a redshift, z , of 0.158.

Fig. 4.8 Satellite-based coordinate systems. *Image* Malcolm Macdonald

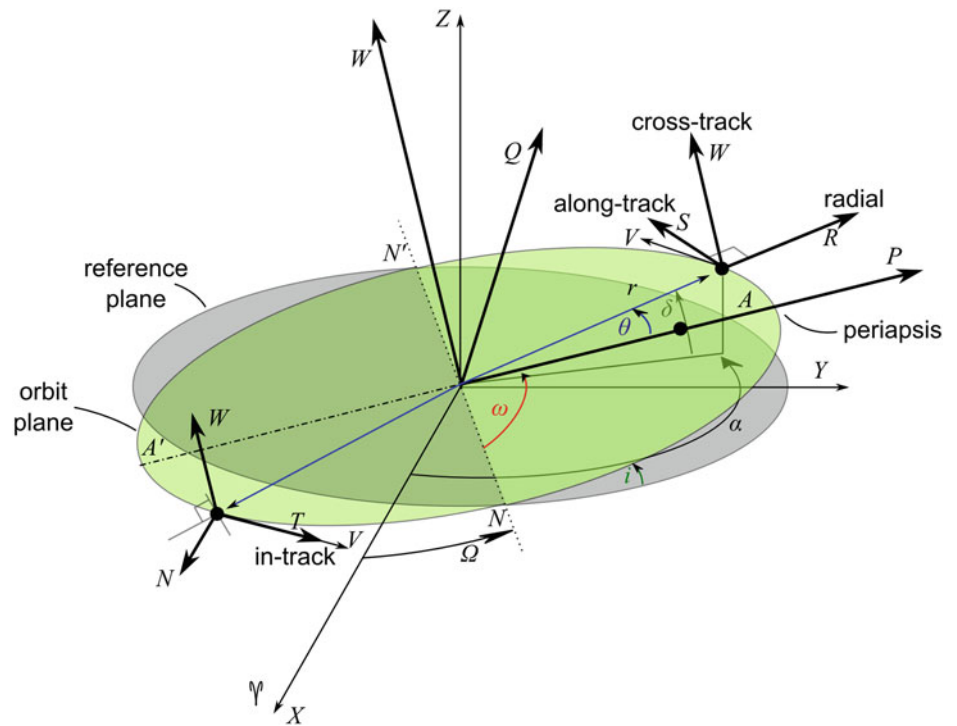
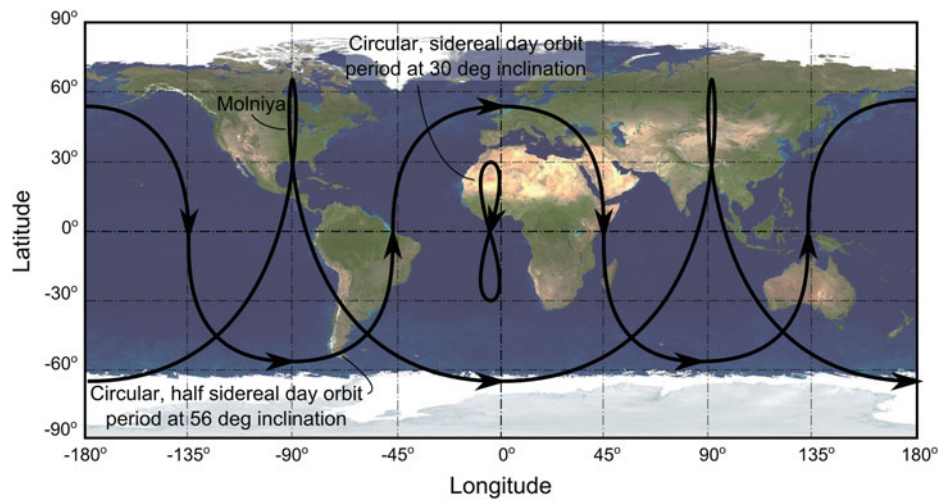


Fig. 4.9 Ground tracks of three different orbits. *Image* Malcolm Macdonald



track to crossover itself and for the orbital motion to appear inverted, as shown in Fig. 4.9 for the Molniya orbit with a period of one half of one sidereal day.

As the semi-major axis of an orbit, and hence its period, is increased towards one sidereal day the apparent rotation of the ground track will compress longitudinally until the ground track appears to repeat over a single portion of the Earth, when the orbit period is equal to the length of a sidereal day, the orbit is called geosynchronous. For an orbit period greater than one sidereal day, the apparent retrograde motion is stretched by increasing the period.

4.1.6 Time

Time is used to accurately define the instant of an event; the moment is referred to as an epoch, which designates the moment as a date. All time systems count from a given epoch. The internationally accepted civilian calendar is the Gregorian (or Christian) calendar, named after Pope Gregory XIII, which counts years from the birth of Jesus of Nazareth, a central figure of Christianity venerated as the son of God. It should be noted however that many other calendars are also still used throughout the world, for

example in India the national calendar, or Saka calendar, is the official civilian calendar, used alongside the Gregorian calendar.

The *Système international d'unités* or, SI base unit of measurement of time is the second. One second is defined as the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom at rest at a temperature of 0 K. Larger units of time, such as the minute, hour or day are defined on this base unit. However, these are non-SI compliant because they are not decimal and there is no fixed ratio between seconds and these larger units owing to the requirement to occasionally introduce leap second adjustments to the Coordinated Universal Time (UTC) scale in order to keep it close to mean solar time.

4.1.6.1 Sidereal Time

The time of successive crossings of an observer's meridian by an arbitrary fiducial direction in the reference plane is one sidereal day. The local sidereal time, LST, is thus the hour angle of the vernal equinox and depends on the observer's longitude. One 24 h period is 86,400 s and one sidereal day is approximately 86,164.1 s.

4.1.6.2 Mean Solar Time

Earth's orbit about the Sun is an ellipse, thus by Kepler's second law the time of successive crossings of an observer's meridian by the Sun varies throughout the Earth's revolution about the Sun. This variation is further compounded by the fact that from the observer's location the path of the Sun is in the ecliptic plane, which is not coincident with the equatorial plane. To overcome this difficulty a fictitious mean Sun is introduced that has nearly uniform motion along the celestial equator. The mean Sun increases its hour angle by 24 h in a sidereal day and thus the time interval of successive crossings of an observer's meridian by the Mean Sun is constant and termed a mean solar day. One mean solar day is equal to approximately 86,400 s, which is less than one sidereal day; however, the relationship is not a constant ratio over a period of centuries due to the non-uniform rotation of the Earth.

The mean solar time is determined by measurements of the Earth's orientation. The mean solar time at Greenwich is termed universal time (UT) and is distinct from Greenwich Mean Time (GMT) which is the apparent local solar time at Greenwich. Universal time is found by reducing the observations of radio galaxies such as quasars from many observation locations, this is termed UT0. UT0 is then corrected for polar motion such that time becomes independent of the observer's location; this time is termed UT1.

4.1.6.3 International Atomic Time

Officially introduced at the start of 1972, International Atomic Time (TAI, from the French name Temps Atomique International) is a high-precision time standard based on the SI definition of a second; that is, by counting the transitions of caesium-133 between two hyperfine levels. However, relativistic effects, such as location, affect the rate of atomic transitions. As such, TAI is determined as a weighted average, including known correction factors, of over 200 atomic clocks from around the world to provide a unit of time that is as close to the SI second as reasonably possible. As such, TAI provides a measure of time independent of the motion of the Earth or Sun.

4.1.6.4 Coordinated Universal Time

The most commonly used time, Coordinated Universal Time (UTC, also referred to as Zulu time), is the principal time standard by which world time is coordinated. UTC is based on TAI, with leap seconds added at irregular intervals in order to maintain UTC within ± 0.9 s of UT1. When introduced in 1972, TAI was offset from UTC by 10 s. By the end of 2013, 25 leap seconds had been introduced and hence TAI was offset from UTC by 35 s. For civilian purposes, local time is defined as offsets from UTC, creating the time zones that are used today throughout the world.

It should be noted that the use of leap seconds is a matter of ongoing debate. In July 2005 the US proposed to the International Earth Rotation and Reference Systems Service (IERS) to eliminate leap seconds from the UTC standard maintained by the ITU Radiocommunication Sector (ITU-R), part of the International Telecommunication Union (ITU). Discussion and resolution of this proposal has been postponed several times due to the controversial nature of the proposal to decouple civilian time from solar time. Resolution is due at the 2015 World Radio Conference.

4.1.6.5 Julian Date

In the observation of a body, it is convenient to state the moment of the observation as a decimal number of days from a given epoch. The Julian Date, JD, is the time, measured in days, from the epoch January 1, 4713 BC, 1200 h (UT1); the reference epoch of the Julian period, a chronological interval of 7,980 years. The next Julian Period begins in the year 3268 AD. Note that the epoch is counted from in the proleptic Julian calendar until October 4, 1582, the last day of the Julian calendar, and thereafter the Gregorian calendar, which started the next day as October 15, 1582 to account for the accumulated drift of the seasons through the year over the preceding thirteen centuries. Note that leap seconds are typically excluded from Julian Day calculations as they are not predictable with a simple formula. The JD epoch is midday such that

astronomical observations, taken at night, can be made in a single Julian Day.

The value of JD is typically very large. Hence, the International Astronomical Union recommends the Modified Julian Date (MJD), be used. The MDJ is taken from an epoch of November 17, 1858, 0000 h (UT1); note that MJD starts at 0000 h, UT1, rather than 1,200 h. An alternatively used epoch is January 01, 2000, 1,200 h (UT1), and is termed J2000.

4.1.7 The Three-Body Problem

The many or n -body problem was first formulated by Newton and seeks, given at any time the positions and velocities of three or more massive particles, the mass of each being known and which are moving under their mutual gravitational force alone, to determine the positions and velocities at any other time. The complexity of this problem has motivated study by many minds over the last three centuries. It is probable that no general solution to this problem is possible; yet the problem can be further complicated by taking into account the shape and internal composition of each body. Note however that several general and useful assertions can be made on the n -body problem; these assertions are expressed in the ten known integrals of motion, see [3, 4] for further information on these.

4.1.7.1 Circular Restricted Three-Body Problem

The Circular Restricted Three-Body Problem (CRTBP), reduces the general three-body problem through two principal assumptions, specifically that two massive particles move in circles about their center of mass, while attracting a third infinitesimal mass to which they are not attracted. The orbits and masses of each of the massive particles being known, the problem is reduced to determining the extent of the possible motion of the third particle. This simplification significantly reduces the order of the problem, whilst providing a good approximation to many problems within astrodynamics, such as the motion of a spacecraft in the vicinity of both the Earth and the Moon. It should be noted that although the assumption of circular motion by the massive particles can be removed, to form the Elliptic Restricted Three-Body Problem (ERTBP), this problem is significantly more complex and often not required within flight dynamics.

Using a synodic reference system the CRTBP is illustrated in Fig. 4.10. The unit of distance between the primary particles, denoted R , is chosen to be one. Similarly, the unit of time is chosen such that the gravitational constant, G , is also one and the total mass of the system, M , is set equal to one. From Fig. 4.10 it is seen that the system rotates about

its center of mass, M , with an angular velocity $\omega = \sqrt{GM/R^3} = 1$. The total mass of the system is $M = (m_1 + m_2)$, whilst the mass ratio $\mu = m_2/(m_1 + m_2)$, where $\mu \leq 1/2$ and $m_1 (= (1 - \mu)M)$ is generally greater than $m_2 (= \mu M)$. Note that as such m_1 is located at $-\mu$ on the X -axis, while m_2 is located at $(1 - \mu)$ on the X -axis. The equations of motion of the third particle of infinitesimal mass can thereafter be derived as

$$\ddot{x} - 2\dot{y} - x = -(1 - \mu) \frac{(x + \mu)}{r_1^3} - \mu \frac{(x - (1 - \mu))}{r_2^3} \quad (4.25)$$

$$\ddot{y} - 2\dot{x} - y = -\left(\frac{1 - \mu}{r_1^3} + \frac{\mu}{r_2^3}\right)y \quad (4.26)$$

$$\ddot{z} = -\left(\frac{1 - \mu}{r_1^3} + \frac{\mu}{r_2^3}\right)z \quad (4.27)$$

where, $r_1 = \sqrt{(x + \mu)^2 + y^2 + z^2}$ and $r_2 = \sqrt{(x - (1 - \mu))^2 + y^2 + z^2}$. Note that in some literature the X -axis is defined positive towards the larger of the two massive bodies and as such the sign of the locations of m_1 and m_2 on the X -axis can be inverted with a corresponding change of signs in Eq. 4.25.

4.1.7.2 Jacobi Integral

Defining the 3-body potential as

$$U = \frac{1}{2}(x^2 + y^2) + \frac{1 - \mu}{r_1} + \frac{\mu}{r_2} \quad (4.28)$$

and determining $\partial U/\partial x$, $\partial r_1/\partial x$, $\partial r_2/\partial x$, $\partial U/\partial y$, $\partial r_1/\partial y$, $\partial r_2/\partial y$, $\partial U/\partial z$, $\partial r_1/\partial z$, $\partial r_2/\partial z$, Eqs. 4.25–4.27 become

$$\ddot{x} - 2\dot{y} = \frac{\partial U}{\partial x} \quad (4.29)$$

$$\ddot{y} - 2\dot{x} = \frac{\partial U}{\partial y} \quad (4.30)$$

$$\ddot{z} = \frac{\partial U}{\partial z}. \quad (4.31)$$

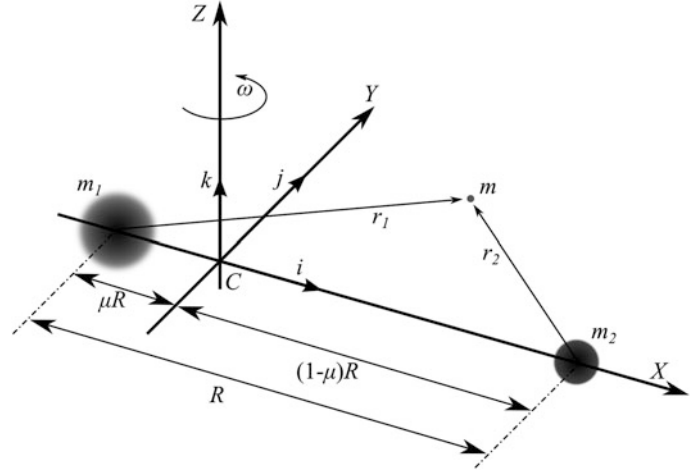
Multiplying by \dot{x} , \dot{y} , \dot{z} respectively and adding together produces a perfect differential that can be integrated to obtain Jacobi's integral as

$$\dot{x}^2 + \dot{y}^2 + \dot{z}^2 = V^2 = 2U - C \quad (4.32)$$

where C is a constant of integration and is termed the Jacobi constant. Jacobi's integral is the only one that can be obtained in the CRTBP.

Using Jacobi's integral the motion of the third particle of infinitesimal mass can be bound. If $2U > C$ then $V^2 > 0$ and

Fig. 4.10 Circular restricted three-body problem geometry. *Image* Malcolm Macdonald



motion is possible, but if $2U < C$ then $V^2 < 0$ and motion is not possible. The boundary between the allowed and forbidden spaces is called Hills limiting surface, a surface of zero velocity, occurring when $2U = C$, or

$$v^2 = (x^2 + y^2) + \frac{2(1-\mu)}{r_1} + \frac{2\mu}{r_2} - C. \quad (4.33)$$

Hills limiting surface is a three-dimensional surface, but by setting $z = 0$ a curve can be found in the plane of motion of the two massive particles. Thereafter for $v = 0$, if r_1 and r_2 are large then the reciprocal of r_1 and r_2 is small and $C \sim (x^2 + y^2)$, the equation of a circle. However, if r_1 and r_2 are small then the reciprocal of r_1 and r_2 is large and $C \sim 2(1-\mu)/r_1 + 2\mu/r_2$, giving small oval curves about m_1 and m_2 . As the Jacobi constant is a function of the initial position and velocity of the infinitesimal mass it is seen that as initial velocity is increased, the Jacobi constant will decrease, allowing the infinitesimal mass to access larger regions of space until eventually it can cross from m_1 to m_2 , and then escape from m_1 and m_2 .

4.1.7.3 Lagrange Points

Although no closed-form solution of Eqs. 4.25–4.27 exists, five equilibrium points, termed Lagrange points, can be derived using these equations together with the Jacobi integral. Specifically, the Lagrange points are double points where the partial derivative of the integral disappears. Joseph-Louis Lagrange [born Giuseppe Luigi Lagrangia, 1736–1813] showed that the required conditions for these equilibrium points are

1. The resultant force on each mass passes through the center of mass of the system.
2. This resultant force is directly proportional to the distance of each mass from the center of mass.

3. The initial velocity vectors are proportional in magnitude to the respective distances of the particles from the center of mass, and make equal angles with the radius vectors to the particles from the center of mass.

Mathematically, in equilibrium the potential defined in Eq. 4.28 must equal zero. Thus considering

$$\frac{\partial U}{\partial x} = x - (1-\mu)\frac{(x+\mu)}{r_1^3} - \mu\frac{(x-(1-\mu))}{r_2^3} = 0 \quad (4.34)$$

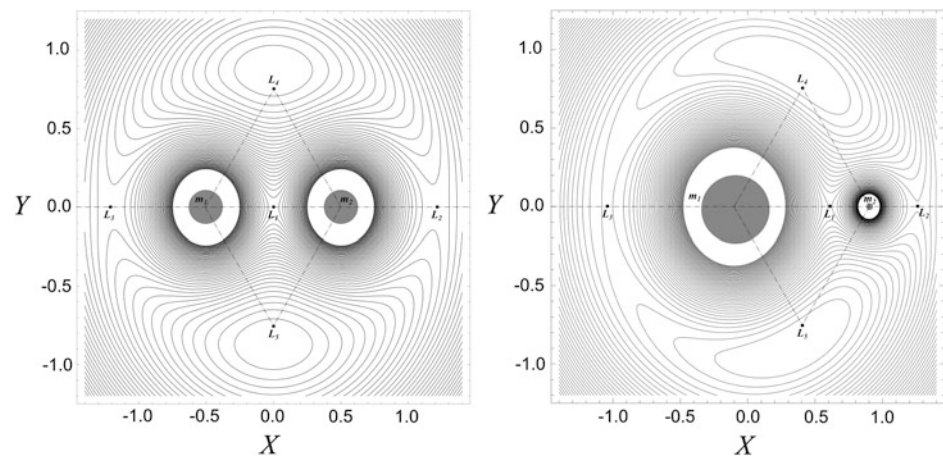
$$\frac{\partial U}{\partial y} = \left(1 - \frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\right)y = 0 \quad (4.35)$$

$$\frac{\partial U}{\partial z} = -\left(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\right)z = 0 \quad (4.36)$$

it is immediately apparent that for equilibrium solutions z must equal zero; in other words, the Lagrange points exist only within the plane of motion of the two massive particles. Similarly, if $\partial U/\partial x = \partial U/\partial y = 0$, then r_1 and r_2 must both equal one. This defines the two equilateral equilibrium points, L_4 and L_5 , both of which can be shown to be stable. In addition to the stable L_4 and L_5 points, three collinear saddle points, L_1 , L_2 and L_3 , can be found on the X -axis; see [3, 4] for a full derivation of these.

Contours of Jacobi constant, C , for two systems with different mass ratio are shown in Fig. 4.11. Note when $\mu = 0.5$, the limiting case, the contours are completely symmetrical, whilst when $\mu = 0.1$ the contours are distorted to reflect the relative mass distribution of the system. Figure 4.11 also shows the locations of each of the Lagrange points. Note how the contours of Jacobi constant are distorted, whilst the L_4 and L_5 points remain equilateral. The contours of Jacobi constant also give a clear illustration of the saddle equilibrium points L_1 , L_2 and L_3 .

Fig. 4.11 Contours of Jacobi constant with Lagrange points illustrated for mass ratios, μ , of 0.5 (left) and 0.1 (right); units are non-dimensional. *Image* Malcolm Macdonald



4.2 Perturbation Techniques

Motion described by laws such as those of Kepler and Newton are idealized approximations of truth that is based on macroscopic observations, such as those of Tycho Brahe. As already stated, the term Keplerian motion is used to describe motion that exactly satisfies Kepler's laws, but such motion is (virtually) impossible to observe in natural celestial bodies due to perturbing forces. That is, microscopic observations show deviations from the macroscopic; perturbation techniques are used to understand such deviations as well as to allow accurate predictions for the motion of a celestial body or a spacecraft.

A notable historic example of the use perturbation techniques was the work in the second quarter of the 19th century by European astronomers to locate the then undiscovered planet Neptune. Based on observed deviations in the orbit of Uranus from published astronomical tables it was hypothesized that an unknown body was causing a gravitational perturbation on the orbit of Uranus. Based on this hypothesis John Couch Adams (1819–1892) and Urbain Le Verrier (1811–1877) developed calculations as to the likely location of the body that was causing the gravitational perturbation. Neither Adams nor Le Verrier were able to muster significant interest from the astronomical community for these predictions. However, following a letter from Le Verrier to Berlin Observatory, Neptune was formally discovered on September 23, 1846 within 1° of where Le Verrier had predicted. It is interesting to note that Cambridge Observatory, on the request of Adams, had actually observed and recorded Neptune twice in the month preceding its formal discovery, yet due to an overly casual approach had failed to recognize what was being observed and hence failed to make the formal discovery. It is also of note that Galileo Galilei (1564–1642) observed and recorded the position of Neptune in December 1612 and January 1613, but as it had just turned 'apparent retrograde' its

motion was only very slight against the celestial sphere and hence not noted by Galileo.

Perturbation techniques are subdivided into two branches termed 'special perturbations', the primary focus of this section, and 'general perturbations'. General perturbations techniques seek to solve the general differential equations of motion, for a certain scenario, in algebraic and/or trigonometric form over a certain time interval and any variation from this is assumed to be slow. Conversely, special perturbations techniques solve the equations of motion, including all necessary perturbations, using direct numerical integration and are, in theory, not time limited. Note that results from general perturbations are often incorporated into special perturbations solutions to account for specific disturbing forces.

Special perturbation techniques allow the state vector, or element set, of a body at a given epoch to be propagated over a short time interval, accounting for all the forces on the body during this interval, using the equations of motion. This calculation can be performed by a variety of methods, enabling the new positions and velocities at the end of the time interval to be found. A second computation using the new positions and velocities enables the process to be carried forward through another time interval. Each computation is called a step and, in theory, the numerical integration can be continued as long as desired. However, in reality rounding errors are introduced and the accuracy of any calculation decreases with every step. A potential (partial) solution to this error is to work with more significant figures than required, such that the final rounding error does not influence the calculation when rounded to the required number of significant figures. Additionally, the error can be further alleviated by the use of as large a time step as possible during each calculation step, thus minimizing the number of occasions on which the solution is rounded. This error source hints at a further drawback of special perturbations techniques, specifically that the state vector must be determined at multiple locations along the trajectory even if only the final condition is desired.

4.2.1 Cowell's Method

Perhaps the most straightforward method of determining the position and velocity of a body is to directly integrate the equations of motion in rectangular coordinates as first performed for a space body in 1908 by Philip Herbett Cowell (1870–1949) and Andrew Clause de la Cherois Crommelin (1865–1935) [6, 7]. The integration formulas used by Cowell and Crommelin were actually first given by Carl Friedrich Gauss (1777–1855). Cowell and Crommelin formulated their equations in rectangular coordinates using Eq. 4.43. Specifically

$$\ddot{\mathbf{r}} + \frac{\mu}{r^3} \mathbf{r} = \mathbf{F} \quad (4.37)$$

where \mathbf{F} is a disturbing acceleration. Note that this can also be written as

$$\ddot{\mathbf{r}} = \nabla(U + F) \quad (4.38)$$

where F is a disturbing potential. The equations of motion are then integrated numerically by means of a multi-step algorithm. Since the publication of the paper by Cowell and Crommelin the use of the term Cowell's method has become ambiguous. In numerical analysis texts 'Cowell-type methods' refer to multi-step algorithms similar to those used in the original paper [3]. However, in celestial mechanics the term 'Cowell's method' often refers to the formulation of the equations in a rectangular coordinate system and the subsequent integration using any technique whatsoever, for example by Runge–Kutta formulas [3, 8, 9]. This would perhaps more correctly be termed 'Cowell's formulation'. Such a method is good for scenarios where the disturbing force or acceleration is of the same or higher order as that due to the central body, as the method does not distinguish between the two. This however is also the primary disadvantage, because a large number of significant figures have to be carried due to the large central force term, requiring many more time steps when the disturbing force or acceleration is small, otherwise a significant loss of accuracy occurs.

4.2.2 Encke's Method

If only the differential accelerations are integrated, rather than the total acceleration, considerable accuracy can be obtained with a larger time interval when the disturbing force or acceleration is small. This method is known as Encke's Method, after Johann Franz Encke (1791–1865) but it was actually first proposed by George Phillips Bond (1825–1865) and William Cranch Bond (1789–1859) of Harvard University in 1849, 2 years before Encke's work became known [3, 8, 9]. As already shown, to a first

approximation an orbit is a conic section, this assumption is at the nucleus of Encke's method. Integrating the difference between the primary acceleration and the perturbing acceleration implies a reference orbit must be employed, along which the body would move in the absence of any perturbations. The integration gives the difference between the real coordinates and the conic section coordinates. The conic section orbit is an osculating orbit, thus at the epoch of osculation the differences vanish. As time from the initial epoch increases so the difference between the real coordinates and the conic section coordinates increases, until it becomes necessary to derive a new osculating orbit. If a new osculating orbit is not derived the various accelerations will grow in magnitude and the process becomes cumbersome. The process of selecting a new conic section from which to calculate deviations is called '*rectification of the orbit*'. Following rectification of the orbit the initial conditions for the deviation vector differential equation are again zero and the only non-zero acceleration is the disturbing acceleration. The error in determining the position and velocity of the osculating orbit is subject only to round-off errors and is independent of the integration technique used. The accuracy of calculation of the deviation from the osculating orbit is limited by both round-off and truncation errors. The integrated quantities are small with respect to the osculating quantities and have little effect on the determination of the true orbit because before the errors become significant a new osculating orbit is selected through the process of rectification. The main advantage of Encke's method is the larger integration intervals that can be adopted compared to Cowell's method. However, the computational cost of a single Encke integration step is much greater than that of a Cowell step. The greater computational cost per step is typically more than compensated for by the larger step size. Encke's method has many applications, for example orbit determination of highly eccentric comets, such as the analysis performed by Encke on a comet later named after him. The method can also be used to analyze orbits in Earth–Moon space, where the Moon is taken as a perturbing body.

It has been shown that in propagating a near-Earth satellite the inclusion of the first-order effects of Earth oblateness, see Sect. 4.3, in the reference orbit greatly improves Encke's method by increasing both the interval between rectifications of the reference orbit and the accuracy of the integration compared with the classic form of Encke's method [10]. It has also been shown that the calculation time for the integration of the motion of four or more bodies can be reduced by an order of magnitude by comparison to the original Encke method if the reference orbit is taken to be a combination of several Keplerian orbits [11]. It is thus clear that the Encke method is

optimized when the reference orbit is known and remains very close to the real evolving orbit for a significant period.

4.2.3 Variation of Parameters

Initially, variation of parameters may appear more problematical to implement than Encke's method, however it has some advantages when the perturbing acceleration is quite small. One of the primary differences is that the Encke reference orbit is constant until rectification occurs, whereas in variation of parameters the reference orbit is continuously changing and may thus be regarded as a form of Encke's method.

The variation of parameters equations of motion are a system of first-order differential equations that describe the rate of change for the time-varying elements. It is from this that the method of variation of parameters gets its alternative name of 'variation of orbital elements', or the slightly paradoxical 'variation of constants'. In 1782, Lagrange completely developed for the first time the method of variation of parameters while studying the elliptical motion of comets. In doing so, Lagrange developed the variational equations of the motion of the classical orbit elements illustrated in Fig. 4.8. The equations of motion are termed Lagrange's planetary equations, the derivation of which can be widely found within the literature; see, for example, [3, 4, 9].

Lagrange's variational equations can either be derived for the special case in which the disturbing acceleration is represented as the gradient of the disturbing function, or they can be derived appropriate to the various choices of component resolutions of the disturbing acceleration vector in the Gaussian, or RTN satellite-based coordinate system; illustrated in Fig. 4.8 as axis *RSW*. This form of the equations of motion is attributed to Gauss

$$\frac{da}{dt} = \frac{2a^2}{\sqrt{\mu p}} [R \quad T \quad N] \begin{bmatrix} e \sin \theta \\ (1 + e \cos \theta) \\ 0 \end{bmatrix} \quad (4.39)$$

$$\frac{de}{dt} = \sqrt{\frac{p}{\mu}} [R \quad T \quad N] \begin{bmatrix} \sin \theta \\ \cos \theta + \cos E \\ 0 \end{bmatrix} \quad (4.40)$$

$$\frac{di}{dt} = \frac{r}{\sqrt{\mu p}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ \cos(\theta + \omega) \end{bmatrix} \quad (4.41)$$

$$\frac{d\Omega}{dt} = \frac{r}{\sqrt{\mu p}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ \left(\frac{\sin(\theta + \omega)}{\sin i} \right) \end{bmatrix} \quad (4.42)$$

$$\frac{d\omega}{dt} = \frac{1}{\sqrt{\mu}} [R \quad T \quad N] \begin{bmatrix} -\frac{\sqrt{p}}{e} \cos \theta \\ \left(1 + \frac{r}{p}\right) \frac{\sqrt{p}}{e} \sin \theta \\ -\frac{r}{\sqrt{p}} \cot i \sin(\theta + \omega) \end{bmatrix}. \quad (4.43)$$

Equations 4.39–4.43 can be used to propagate a trajectory with the inclusion of a sixth position-fixing element. The sixth element could be, amongst other things, the eccentric anomaly, mean anomaly, true longitude, or as shown in Eq. 4.44, the true anomaly

$$\frac{d\theta}{dt} = \frac{\sqrt{\mu p}}{r^2} \frac{ep\mu}{ep\mu + r^2(-pR \cos \theta + (p+r)T \sin \theta)} \quad (4.44)$$

which reduces to Eq. 4.11 in the absence of any perturbing forces.

Lagrange's planetary equations in the Gaussian form can be analytically integrated, as in the method of general perturbations, or they can be integrated numerically step-by-step, with the new elements at the end of each step being used as the basis for the computation of the next step. Since Lagrange first introduced his planetary equations, where the rates of change of the osculating elements of a planet's orbit are given in terms of the elements of that planet and of the planets disturbing its heliocentric orbit, various attempts have been made to overcome some of the serious problems associated with the method. Some of the advantages of the variation of parameters method are that it is strictly a perturbation method and as such bypasses the central-body acceleration. For moderate perturbations, the differentials of the elements are small and as such a larger step size can be used than in a rectangular coordinate method in which the central-body acceleration must be calculated each step. Among the perceived disadvantages of the method is the more complicated nature of the right-hand side of the equations compared to those of the rectangular coordinates equations of motion, including the presence of sine and cosine terms. Additionally, the traditionally perceived disadvantages are the need to solve Kepler's equation, the break-down of the equations when orbit eccentricity is zero or one, or orbit inclination is zero, and the fact that the equations are usually given in elliptical elements and are thus inapplicable to parabolic, hyperbolic or rectilinear orbits. The disadvantages regarding computational difficulties offset some of the benefits of a larger time step than a Cowell type solution. However, such issues can be minimized with modern computing capabilities and prudent programming.

It was noted previously that the classic orbit elements are ill-defined in certain circumstances. This presents several significant difficulties when attempting to propagate a trajectory in these regions using the Gaussian form of Lagrange's planetary equations. For example, as the orbit

eccentricity drops towards zero the rate of change of the apsides becomes indeterminable, see Eq. 4.43. Similarly, as the inclination drops to zero the rate of change of the ascending node becomes indeterminable, see Eq. 4.42. The obvious solution is thus to define the orbit through a change in variables, which can be done simply by using the true longitude or argument of latitude as previously mentioned. Alternatively, this can be done by, for example, noting symmetries to apply standard transformations to make a change of variable from Keplerian to Delaunay variables [12].

4.2.3.1 Non-singular Elements and the Equations of Motion

To derive variational equations that are non-singular, combinations of the classical elements that do not depend on either the line of nodes or the apsidal line are sought. Adding the variational equations for Ω and ω , given in Eqs. 4.42 and 4.43, eliminates the singularity at zero inclination

$$\frac{d\varpi}{dt} = \frac{1}{nabe} [R \quad T \quad N] \begin{bmatrix} -p \cos \theta \\ (p+r) \sin \theta \\ er \sin(\theta + \omega) \tan \frac{i}{2} \end{bmatrix}. \quad (4.45)$$

Noting that

$$\frac{dM}{dt} = n + \frac{1}{a^2 en} [R \quad T \quad N] \begin{bmatrix} p \cos \theta - 2re \\ -(p+r) \sin \theta \\ 0 \end{bmatrix} \quad (4.46)$$

the variational equations for ϖ and M can thus be added to obtain an equation that also removes the singularity due to zero eccentricity

$$\frac{dl}{dt} = n + \frac{1}{n} [R \quad T \quad N] \begin{bmatrix} -\left(\frac{ep \cos \theta}{b(a+b)} + \frac{2r}{a^2}\right) \\ \left(\frac{e(p+r) \sin \theta}{b(a+b)}\right) \\ \left(\frac{r \sin(\omega+\theta) \tan(i/2)}{ab}\right) \end{bmatrix}. \quad (4.47)$$

As Eq. 4.47 is a function of the true anomaly, which is referenced to periapsis, further development is required. Kepler's equation can be written in the augmented form of

$$\begin{aligned} l &= \varpi + M = \varpi + E - e \sin E \\ &= (\varpi + E) + e \sin \varpi \cos(\varpi + E) - e \cos \varpi \sin(\varpi + \theta). \end{aligned} \quad (4.48)$$

Note that the orbit radius may be written as

$$\begin{aligned} r &= a(1 - e \sin \varpi \sin K - e \cos \varpi \cos K) \\ &= \frac{p}{1 + e \sin \varpi \sin L + e \cos \varpi \cos L}. \end{aligned} \quad (4.49)$$

From Kepler's equation, Eq. 4.48, and the equation of an orbit, Eq. 4.49, note that the eccentricity equivalent term and the longitude of periapsis equivalent term only appear in the combinations $e \sin \varpi$ and $e \cos \varpi$. These functions are thus selected to replace e and ϖ respectively. Following a similar process, and writing the argument of latitude in terms of the true longitude, it is possible to select $(\tan(i/2) \sin \Omega)$ and $(\tan(i/2) \cos \Omega)$ to replace Ω and i . This element set is referred to as 'equinoctial elements'. The equinoctial elements are non-singular except for rectilinear orbits and when $i = \pi$. This element set was first introduced by Lagrange in 1774 for his study of secular variations. He used i rather than $i/2$, but the inclusion of the half-angle simplifies the resulting Gaussian equations of motion and allows the use of Allan's expansion of the geopotential, if desired [13].

4.2.3.2 Modified Equinoctial Elements

Employing a 'fast variable' (phase angle) as the sixth or position-fixing element allows a regular perturbation technique to be used, with the fast variable as the independent variable. It thus becomes logical to modify the equinoctial elements by choosing true longitude in place of mean anomaly as the position-fixing element. Furthermore, by replacing the semi-major axis with the semi-latus rectum a set of orbit elements that are non-singular for all orbits excluding $i = \pi$ is obtained; however this singularity can be handled by appropriate definition of a 'retrograde factor'. The 'modified equinoctial elements' are thus defined as

$$p = a(1 - e^2) \quad (4.20)$$

$$f = e \cos(\omega + \Omega) \quad (4.50)$$

$$g = e \sin(\omega + \Omega) \quad (4.51)$$

$$h = \tan \frac{i}{2} \cos \Omega \quad (4.52)$$

$$k = \tan \frac{i}{2} \sin \Omega \quad (4.53)$$

$$L = \Omega + \omega + \theta = \varpi + \theta. \quad (4.54)$$

The auxiliary (positive) variables are

$$s^2 = 1 + h^2 + k^2 \quad (4.55)$$

$$w = 1 + f \cos L + g \sin L \quad (4.56)$$

$$r = \frac{p}{w} \quad (4.57)$$

$$\tau = \sqrt{h^2 + k^2} \quad (4.58)$$

$$\alpha^2 = h^2 - k^2 \quad (4.59)$$

where Eq. 4.57 is simply the orbit radius.

The modified equinoctial elements equations of motion in the Gaussian form are found to reduce to

$$\frac{dp}{dt} = \frac{2p}{w} \sqrt{\frac{p}{\mu}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad (4.60)$$

$$\frac{df}{dt} = \sqrt{\frac{p}{\mu w}} [R \quad T \quad N] \begin{bmatrix} w \sin L \\ (w+1) \cos L + f \\ -(h \sin L - k \cos L)g \end{bmatrix} \quad (4.61)$$

$$\frac{dg}{dt} = \sqrt{\frac{p}{\mu w}} [R \quad T \quad N] \begin{bmatrix} -w \cos L \\ (w+1) \sin L + g \\ (h \sin L - k \cos L)f \end{bmatrix} \quad (4.62)$$

$$\frac{dh}{dt} = \sqrt{\frac{p}{\mu 2w}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ \cos L \end{bmatrix} \quad (4.63)$$

$$\frac{dk}{dt} = \sqrt{\frac{p}{\mu 2w}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ \sin L \end{bmatrix} \quad (4.64)$$

$$\frac{dL}{dt} = \sqrt{\mu p} \left(\frac{w}{p}\right)^2 + \sqrt{\frac{p}{\mu w}} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ (h \sin L - k \cos L) \end{bmatrix}. \quad (4.65)$$

Notice that when the disturbing acceleration is zero Eqs. 4.60–4.64 equal zero, while Eq. 4.65 reduces to the angular momentum term.

4.2.3.3 Transformation from Modified Equinoctial Elements to Classical Elements

The transformation from modified equinoctial elements are obtained as

$$a = \frac{p}{1 - f^2 - g^2} \quad (4.66)$$

$$e = \sqrt{f^2 + g^2} \quad (4.67)$$

$$\begin{aligned} i &= 2 \tan^{-1} \tau = 2 \tan^{-1} \left(\sqrt{h^2 + k^2} \right) \\ &= \tan^{-1} \left(2\sqrt{h^2 + k^2}, 1 - h^2 - k^2 \right) \end{aligned} \quad (4.68)$$

$$\Omega = \tan^{-1} \left(\frac{k}{h} \right) \quad (4.69)$$

$$\tan(\omega + \Omega) = \frac{g}{f} \quad (4.70)$$

$$\omega = \tan^{-1} \left(\frac{g}{f} \right) - \tan^{-1} \left(\frac{k}{h} \right) = \tan^{-1} (gh - fk/fh + gk) \quad (4.71)$$

$$\theta = L - \tan^{-1} \left(\frac{g}{f} \right). \quad (4.72)$$

Using Eqs. 4.66–4.72 the following identities can also be derived

$$\cos \theta = \frac{f \cos L + g \sin L}{\sqrt{f^2 + g^2}} \quad (4.73)$$

$$\sin \theta = \frac{f \sin L - g \cos L}{\sqrt{f^2 + g^2}} \quad (4.74)$$

$$\cos \Omega = \frac{h}{\tau} \quad (4.75)$$

$$\sin \Omega = \frac{k}{\tau} \quad (4.76)$$

$$\cos i = \frac{1 - \tau^2}{1 + \tau^2} \quad (4.77)$$

$$\sin i = \frac{2\tau}{1 + \tau^2} \quad (4.78)$$

$$\cos \omega = \frac{fh + gk}{\tau \sqrt{f^2 + g^2}} \quad (4.79)$$

$$\sin \omega = \frac{gh - fk}{\tau \sqrt{f^2 + g^2}} \quad (4.80)$$

$$\cos(\omega + \theta) = \frac{h \cos L + k \sin L}{\tau} \quad (4.81)$$

$$\sin(\omega + \theta) = \frac{h \sin L - k \cos L}{\tau}. \quad (4.82)$$

4.2.3.4 Transformation from Modified Equinoctial Elements to Cartesian Form

The relationship between the modified equinoctial element set and the state vector, that is the position and velocity vectors, is

$$\mathbf{r} = \frac{r}{s^2} \begin{bmatrix} \cos L + \alpha^2 \cos L + 2hk \sin L \\ \sin L - \alpha^2 \sin L + 2hk \cos L \\ 2(h \sin L - k \cos L) \end{bmatrix} \quad (4.83)$$

$$\mathbf{v} = \frac{1}{s^2} \sqrt{\frac{\mu}{p}} \begin{bmatrix} \alpha^2 \sin L + \sin L - 2hk \cos L + g - 2fhk + \alpha^2 g \\ \alpha^2 \cos L - \cos L + 2hk \sin L - f + 2ghk + \alpha^2 f \\ -2(h \cos L + k \sin L + fh + gk) \end{bmatrix}. \quad (4.84)$$

Note that it is also possible to compute the inverse transformation. However, the true longitude can only be defined to within a multiple of 2π and thus the reference epoch must be known in order to resolve its actual value.

4.2.4 Numerical Integration

Numerical integration methods can be divided into either the single-step or multi-step. The difference between these two methods is well illustrated in [3]. However, the difference can be summarized by noting that a single-step method is a self-starting method that only uses data from the beginning of the current step in the calculation of the variable values at the end of the step. Furthermore, changing the step-size to match a defined error criterion poses no difficulties, easily allowing the interval step-size to be halved or doubled. The primary difficulty with a single-step method is that if the equations are non-linear, such as Lagrange's planetary equation of motion, then it may become a time-consuming and unwieldy process to calculate the higher-order terms of the expansion. A multi-step method allows larger interval step-sizes to be adopted even when the higher-order terms of the expansion are calculated. However, the law of diminishing returns sets in. Furthermore, stability considerations mean that it is wise to keep the order below double figures. A multi-step procedure involves fewer computations than a single-step method, correct to the same order, subject to the constraint of not being self-starting and that special procedures are required to half or double the step-size. Therefore, multi-step methods are best suited to scenarios where the step-size changes can be removed or minimized, such as almost circular orbits, or when the equations have been regularized.

4.2.4.1 Errors

It can be shown that the probable error of a double integral is $0.1124n^{3/2}$, where n is the number of integration steps [14]. That is to say, after numerically integrating the second-order (x , y , z) equations of motion, Eq. 4.37, through 100 steps there is an even odds chance that the rounding error is smaller than 112.4 in units of the last decimal [3]. Furthermore, the mean error of the osculating elements of a body obtained by numerically integrating the Lagrange planetary equations, which are first-order, will be proportional to $n^{1/2}$, apart from the mean orbital longitude (or whatever position fixing element is selected) in which case the mean error is again proportional to $n^{3/2}$ as this is a result of a double integral [14].

4.3 Disturbing Force

The nature of the disturbing force that acts to cause deviation from the macroscopic truths observed by Kepler in Tycho Brahe's observations is varied and diverse. It should also be noted that deviation from the laws of motion derived by Kepler and Newton cannot at all times be ignored in the preliminary mission design stage as this deviation may actually be key to enabling the mission concept, as will be discussed in Sect. 4.4.

4.3.1 Additional Gravitational Fields

Incorporating the effects of the gravitational attraction of a massive particle, that is, other than the primary body, can be achieved by formulating the problem in rectangular coordinates. In doing so the force \mathbf{F} of n additional massive particles on the massless particle, that is, a spacecraft located at radius \mathbf{r} , can be expressed as

$$\mathbf{F} = G \sum_j^n m_j \left(\frac{\mathbf{r}_j - \mathbf{r}}{|\mathbf{r}_j - \mathbf{r}|^3} - \frac{\mathbf{r}_j}{|\mathbf{r}_j|^3} \right) \quad (4.85)$$

and thereafter incorporated into Eq. 4.37 for the numerical propagation of the spacecraft's state vector. Note that if the Gaussian form of Lagrange's variational equations is used for the numerical propagation of the spacecraft's element set then the disturbing force given by Eq. 4.85 must be converted into the satellite-based Gaussian or RTN coordinate system.

For spacecraft in Earth orbit, the principal gravitational perturbations are due to the Moon and the Sun, whilst Jupiter is the most significant of the planets. However, the disturbing force due to Jupiter's gravitational acceleration is typically at least five orders of magnitude less than either the Moon or Sun. It can also be shown that the disturbing force of the Sun on a spacecraft in orbit about the Moon is greater than that on one in orbit about the Earth. Finally, it should be noted that typically the orbit of the disturbing body will not be in the same plane as that of the massless particle, that is, the spacecraft, hence the disturbing force will act to cause a change in the orbit plane orientation.

4.3.1.1 Sphere of Influence

Recognizing that the two-body problem is perturbed by additional gravitational fields, the concept a sphere of influence defines an almost spherical region, centered about the primary body, within which the motion of a massless particle, that is, the spacecraft, can be considered as primarily determined by the gravitational attraction of the

primary body. Using Eqs. 4.37 and 4.85 the equations of motion of a massless particle, C , with respect to two massive particles, A and B , in rectangular coordinates may be written as

$$\ddot{\mathbf{r}}_{AC} + \frac{\mu_A}{r_{AC}^3} \mathbf{r}_{AC} = -Gm_B \left(\frac{\mathbf{r}_{BC}}{r_{BC}^3} + \frac{\mathbf{r}_{AB}}{r_{AB}^3} \right) \quad (4.86)$$

$$\ddot{\mathbf{r}}_{BC} + \frac{\mu_B}{r_{BC}^3} \mathbf{r}_{BC} = -Gm_A \left(\frac{\mathbf{r}_{AC}}{r_{AC}^3} - \frac{\mathbf{r}_{AB}}{r_{AB}^3} \right) \quad (4.87)$$

where subscript A and B denote the massive particle and radius subscripts AB , BC and AC denote vector directions between particles. The equations of motion can thus be written as

$$\ddot{\mathbf{r}}_{AC} + \mathbf{Y}_A = \Phi_B \quad (4.88)$$

$$\ddot{\mathbf{r}}_{BC} + \mathbf{Y}_B = \Phi_A. \quad (4.89)$$

The \mathbf{Y}/Φ ratios thus give the order of magnitude of the perturbation within each system due to the other massive particle. The sphere of influence is where the ratios are equal.

In practice if a spacecraft, particle C , is in orbit about a planet, particle B , which is in turn in orbit about a star, particle A , or similarly if particle B is a moon in orbit about a planet, then $\mathbf{r}_{AC} \gg \mathbf{r}_{BC}$. Hence, François Félix Tisserand (1845–1896) showed that the surface defined as the sphere of influence of particle B is almost spherical with radius

$$r_{SoI.A} = \left(\frac{m_B}{m_A} \right)^{2/5} r_{AB}. \quad (4.90)$$

Equation 4.90 defines the sphere of influence as a single surface. However, practically it is clear that the influence of an additional gravitational field is not realized at a boundary. It is therefore useful in astrodynamics to consider the development of two spheres of influence. If the influence of the massive particle B can be neglected when the perturbation on the massless particle, C , is less than a given fraction, ϵ_B , of the acceleration due to massive particle A then

$$|\mathbf{Y}_A| = \epsilon_B |\Phi_B| \quad (4.91)$$

defines an outer sphere of influence beyond which motion can be described by the two-body problem about the massive particle A . Similarly

$$|\mathbf{Y}_B| = \epsilon_A |\Phi_A| \quad (4.92)$$

describes an inner sphere of influence within which perturbation due to massive particle A can be neglected and motion can be described by the two-body problem about the massive particle B . Hence, between the two spheres

prolonged periods of motion cannot be accurately described by the two-body problem. Defining the mass ratio of the two massive particles as $m_* = m_B/m_A$, and the radius ratio of the massless particle from the two massive particles as $r_* = r_B/r_A$ then

$$|\epsilon_B| = \frac{m_*}{r_*^2} \left(1 - \left(\frac{r_*}{1+r_*} \right)^2 \right) \quad (4.93)$$

and

$$|\epsilon_A| = \frac{r_*^2}{m_*} \left(1 - \frac{1}{(1+r_*)^2} \right) \quad (4.94)$$

give values of ϵ_A and ϵ_B for values of r_* .

4.3.2 Non-Spherical Central Body

The uneven distribution of mass within a central body perturbs the gravitational field of the central body from the point-mass representation assumed in the standard two-body problem. Taking the equations of motion written in the potential form, see Eq. 4.38, it is convenient to describe the gravitational field of a non-spherical central body, outwith its surface, using a spherical harmonic expansion. Hence, assuming the origin of the reference frame is coincident with the central body's center of mass, this may be written as

$$U = \frac{\mu}{r} \left[1 + \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{R}{r} \right)^n P_{nm}(\sin \phi) (C_{nm} \cos(m\lambda) + S_{nm} \sin(m\lambda)) \right] \quad (4.95)$$

where R is the mean equatorial radius of the central body, ϕ is the geocentric latitude, ϕ^l in Fig. 4.7, of the sub-point of the massless body on the reference ellipsoid, and λ is the equivalent longitude, recalling that longitude is always geocentric. C_{nm} and S_{nm} are the gravitational coefficients dependent on the mass distribution of the central body; when $n \neq m$ these are tesseral harmonic coefficients and when $n = m$ they are sectoral (or sectorial) harmonic coefficients. Note that the sectoral harmonic coefficients, when $n = m$, represent bands of longitude, dividing the sphere into $2n$ longitudinal bands. Whilst the tesseral harmonic coefficients, when $n \neq m$, as the name suggests² modify the gravitational potential to model specific regions as 'tiles' on the reference ellipsoid. Finally, note that the Legendre polynomials in Eq. 4.94, P_{nm} , have the form

² A tessera is an individual tile in a mosaic.

$$P_{nm}(\sin \phi) = (\cos^2 \phi)^{\frac{m}{2}} \frac{d^m}{d(\sin \phi)^m} P_n(\sin \phi) \quad (4.96)$$

$$P_n(\sin \phi) = \frac{1}{2^n n!} \frac{d^n}{d(\sin \phi)^n} (\sin^2 \phi - 1)^n. \quad (4.97)$$

It is common to write Eq. 4.94 using a J_n notation for the zonal harmonic gravitational coefficients, as $J_n = -C_{n,0}$, and separate these terms

$$U = \frac{\mu}{r} \left[1 - \sum_{n=2}^{\infty} J_n \left(\frac{R}{r}\right)^n P_n(\sin \phi) + \sum_{n=2}^{\infty} \sum_{m=1}^n \left(\frac{R}{r}\right)^n P_{nm}(\sin \phi) (C_{nm} \cos(m\lambda) + S_{nm} \sin(m\lambda)) \right]. \quad (4.98)$$

It is seen in Eq. 4.98 that the zonal harmonics, when $m = 0$, remove longitudinal dependencies, making the gravitational field symmetric about the rotational axis of the central body. Zonal harmonics are thus simply longitudinal bands about the central body, such that for any $P_n(\sin \phi)$ there will be n circles of latitude where the Legendre polynomial equals zero and hence $n + 1$ bands where the function oscillates above and below the reference ellipsoid.

It should be noted that if the Gaussian form of Lagrange's variational equations are used for the numerical propagation of the spacecraft's element set that both Eqs. 4.95 and 4.98 include the gravitational acceleration of a point-mass central body. As such, this zero-order effect must be removed before transferring the perturbation vector into the satellite-based Gaussian or RTN coordinate system.

The dominant perturbation due to the Earth's shape is J_2 , which is three orders of magnitude larger than J_3 and dominates the gravitational perturbations at Earth. It is common to approximate the Earth as a body possessing axial symmetry, as the longitudinal variations will typically be balanced over the orbit period as the spacecraft moves around the Earth. It should be noted however that for a spacecraft in geostationary orbit this approximation does not hold, as the spacecraft remains above the same region of the Earth at all times and the spacecraft encounters a form of resonance, often termed a triaxiality, which induces an east–west drift in the spacecraft's position.

4.3.3 Atmospheric Effects

The structure of the atmospheres of the terrestrial planets is discussed in detail in Chap. 5. However, whilst the structure of each atmosphere is distinct, the principal effect they have on an orbiting body is the same; specifically, they act primarily as a retarding force against the velocity vector. The cause of this retarding, or drag force is the particles that

make up the atmosphere. Accurately quantifying the effects of an atmosphere on a spacecraft is immensely difficult, and predicting these effects is even worse. It should be noted that all bodies moving within an atmosphere generate some form of lift force, but for spacecraft this can be neglected for except in the case of high-accuracy analysis. A vast range of factors, each of which can be immensely difficult to quantify in terms of its past behavior, let alone predict into the future, interact in a complex fashion to influence the upper atmosphere, and specifically the density of the particles of which it consists.

As discussed in Chap. 3, atmospheric models can be static profiles, global analytic fits, or time-varying. Static models, while simple, may account for effects including latitudinal variations, where, for example, the Earth's equatorial bulge (which results in the J_2 perturbation) causes a variation in altitude, in turn causing the density of the atmosphere encountered to vary. Similarly, static models may account for longitudinal variations due to other large landmasses, such as the Andes or the Himalayas. Time-varying atmospheric models principally aim to capture causes of temperature fluctuations in the upper atmosphere. Extreme ultraviolet radiation, EUV, from the Sun causes near-instantaneous heating of the upper atmosphere and hence affects atmospheric density. Meanwhile, other causes of atmospheric heating, such as geomagnetic activities, exhibit a cause and effect delay. As such capturing the impact of such variations can be extremely difficult. Some factors which affect the temperature and hence density of the upper atmosphere include

- *Atmospheric rotation*—Atmospheres tend to rotate with the surface of the central body, but shearing effects cause the rate of rotation to decrease at increased altitude.
- *Diurnal variations*—As the planet rotates it exposes different regions of the atmosphere to different levels of solar heating. The warmest region of the atmosphere lags the sub-solar point and occurs at 1,400–1,430 h local time, minimum density is approximately opposite. Note that the time of year is also important here due to the obliquity of the ecliptic, which presents different latitudes to the Sun at different times of the year.
- *Planetary distance variations*—Eccentricities in the orbit of the planet about the Sun cause a change in distance from the Sun to the planet, but these effects are typically small at Earth.
- *Solar cycle*—An approximately 11-year cycle in solar magnetic activity alters the flux of radiation from the Sun at Earth.
- *Solar rotation*—As the Sun rotates different regions of solar activity are directed towards the Earth. Furthermore, the sidereal rotation period at the Sun's equator is

Table 4.2 US standard atmosphere 1976, from [15]

Altitude (km)	Density (kg/m ³)	Pressure (Pa)	Scale height (km)	Molecular weight (kg/kmol)
0	1.225	1.01e+5	8.4345	29.0
150	2.076e−9	4.54e−4	23.380	24.1
200	2.541e−10	8.47e−5	36.183	21.3
250	6.073e−11	2.48e−5	44.924	19.2
300	1.916e−11	8.77e−6	51.193	17.7
350	7.014e−12	3.45e−6	55.832	16.7
400	2.803e−12	1.45e−6	59.678	16.0
450	1.184e−12	6.45e−7	63.644	15.3
500	5.215e−13	3.02e−7	68.785	14.3
550	2.384e−13	1.51e−7	76.427	13.1
600	1.137e−13	8.21e−8	88.244	11.5
650	5.712e−14	4.89e−8	105.992	9.72
700	3.070e−14	3.19e−8	130.630	8.00
750	1.788e−14	2.26e−8	161.074	6.58
800	1.136e−14	1.70e−8	193.862	5.54
850	7.824e−15	1.34e−8	224.737	4.85
900	5.759e−15	1.09e−8	250.894	4.40
950	4.453e−15	8.98e−9	271.754	4.12
1000	3.561e−15	7.51e−9	288.203	3.94

approximately 24.5 days. However, solar rotation varies with latitude as the Sun is composed of gaseous plasma.

4.3.3.1 Atmospheric Density Models

A large number of atmospheric models have been developed for a range of planets, including the range of Global Reference Atmosphere Models, GRAM, for Venus, Earth, Mars, Neptune and Titan based on empirical data and validated against other empirical data and previous models. The GRAM models are well suited to numerical simulation of spacecraft trajectories, using inputs including geographical position, time, solar and geomagnetic data, and data on the upper atmospheric climate, to provide outputs ranging through density, temperature, pressure, winds and atmospheric constituent concentrations. An alternative and convenient Earth atmospheric model for use with analytical methods is the 1976 US Standard Atmosphere [15], which is an ideal, steady-state model of the Earth's atmosphere at a latitude of 45° north during moderate solar activity. A good overview of Earth atmospheric models, and how they have developed, is given in [4]. However, it should be apparent from engineering judgment alone that no single model can be 'best' for all applications. Table 4.2 and Fig. 4.12 give the density, pressure, scale height and molecular weight variation with altitude as defined by the 1976 US Standard Atmosphere [15].

The simplest possible static atmospheric model is developed by considering the gas law, where the temperature, T , pressure, p , and density, ρ , are related by

$$\frac{p}{\rho} = \frac{RT}{M} \quad (4.99)$$

where R is the gas constant (8.3144621 J mol^{−1} k^{−1}) and M is the molecular weight of the gas. From the hydrostatic equation, the decrease of pressure with altitude, y , above the reference ellipsoid is

$$\frac{dp}{dy} = -\rho g \quad (4.100)$$

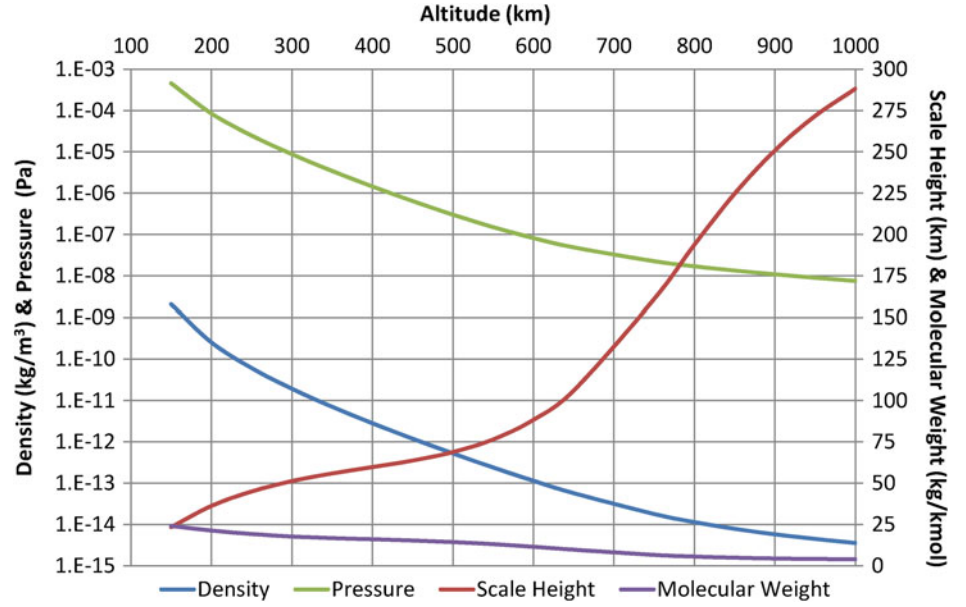
where g is the acceleration due to gravity. Combining Eqs. 4.99 and 4.100, to eliminate ρ , and assuming that (RT/Mg) is a constant, denoted h , the scale height, the variation in pressure with altitude is found as

$$p = p_0 \exp\left(-\frac{y - y_0}{h}\right) \quad (4.101)$$

where p_0 is the pressure at the reference level, when $y = y_0$. Noting that the gas constant can be related to the Boltzmann constant, k_B , using the Avogadro constant, N_A , such that $R = k_B N_A$, the scale height is often also written as

$$h = \frac{k_B T}{M_{mean} g} \quad (4.102)$$

Fig. 4.12 US standard atmosphere 1976, from [15].
Image Malcolm Macdonald



where M_{mean} is the mean molecular mass of dry air in kilograms, which at sea level is 28.97u, or 4.81×10^{-26} kg. Thereafter, assuming that the atmosphere is spherically symmetric, the variation of density with radius r is

$$\rho = \rho_0 \exp\left(-\frac{r-r_0}{H}\right) \quad (4.103)$$

where H is the density scale height and, for a constant h , is

$$\beta = \frac{1}{H} = \left(\frac{1}{h} - \frac{2}{r_0}\right). \quad (4.104)$$

Note that the assumption of a constant scale height requires that the density scale height is also constant. While these equations are not exact, they remain valid to an altitude of several 100 km [16]. The reference altitude, y_0 , can be the top, or bottom, of a specific atmospheric layer, or a nominal altitude, say 100 km, or it may even simply be sea level. However, it is advisable to use a reference height, and hence temperature, as close to the actual height as possible, as (RT/Mg) is not strictly constant; thus, by dividing the atmosphere into thin strips the approximation can be maintained.

4.3.3.2 Equations of Force

It is convenient to represent the retarding force of the atmosphere using the standard drag acceleration equation

$$\mathbf{F}_D = \frac{1}{2} \rho \frac{C_D S}{m} v_r^2 \left(\frac{-\mathbf{v}_r}{|v_r|}\right) \quad (4.105)$$

where \mathbf{v}_r is the velocity relative to the atmosphere, m is the mass of the body, S is the reference surface area of the body,

typically defined as the area of the orthographic projection to a plane perpendicular to the free flow, and C_D is the drag coefficient of the body with respect to the reference surface area. Issues surrounding the drag coefficient will be discussed in more detail in the next section. Note that the ballistic coefficient is usually defined as $(m/C_D S)$, where a low ballistic coefficient means the body is more susceptible to drag forces.

Atmospheric rotation means that the atmosphere will generally induce a force out of the plane of motion, resulting in a change in the orbit plane orientation. Noting that the west-to-east rotation of the atmosphere is of the order 400 m/s and that this is typically greatly in excess of meridional winds (north-to-south, or vice versa). Then to the first-order approximation, where the dominant motion of the atmosphere is west to east, and hence the meridional winds may be neglected. The velocity of the atmosphere in the Earth centered rectangular coordinates system is thus denoted as $\mathbf{v}_a = \mathbf{r} \omega_{atm} \cos \phi$. Thereafter, relating the velocity of the body, \mathbf{v} , to the velocity of the body relative to the atmosphere, \mathbf{v}_r , Eq. 4.105 may be approximated as [16]

$$\mathbf{F}_D = \frac{1}{2} \rho v^2 \frac{C_D S}{m} F \left(\frac{\mathbf{v} - \mathbf{v}_a}{|\mathbf{v} - \mathbf{v}_a|}\right) \quad (4.106)$$

where

$$F = \left(1 - \frac{r_p}{v_p} \omega_{atm} \cos i\right)^2. \quad (4.107)$$

Here ω_{atm} is the mean angular rate of rotation of the atmosphere, which is typically taken to be the same as the

Earth's mean rotation rate but can vary between 0.8 and 1.3 revolutions per day. From Eq. 4.37, the two-body equations of motion thus become

$$\ddot{\mathbf{r}} + \frac{\mu}{r^3}\mathbf{r} = \frac{\rho v^2 C_D S}{2m} F \left(\frac{\mathbf{v} - \mathbf{v}_a}{|\mathbf{v} - \mathbf{v}_a|} \right). \quad (4.108)$$

In a satellite-based Gaussian or RTN coordinate system the disturbing accelerations is [16]

$$R = -\frac{\rho v \delta}{2} \sqrt{\frac{\mu}{pF}} e \sin \theta \quad (4.109)$$

$$T = -\frac{\rho v \delta}{2} \sqrt{\frac{\mu}{pF}} \left(1 + e \cos \theta - r \omega_{atm} \sqrt{\frac{p}{\mu}} \cos i \right) \quad (4.110)$$

$$N = -\frac{\rho v \delta}{2\sqrt{F}} r \omega_{atm} \sin i \cos(\theta + \omega) \quad (4.111)$$

where

$$\delta = (FC_D S/m) \quad (4.112)$$

is a modified ballistic parameter and F is given by Eq. 4.107. An excellent reference for the effects of an atmosphere on an orbiting body is [16] in which various atmospheric phenomena are considered in detail.

4.3.3.3 Drag Coefficient

Determining the drag coefficient of an orbiting body is complicated by the nature of the free molecular flow regime. Free molecular flow occurs when the mean free path, λ , of molecules greatly exceeds a typical linear dimension of a body, l . The ratio λ/l is termed the Knudsen number, Kn , and is discussed in more detail later in this chapter. Free molecular flow applies when Kn is greater than about 10 [16]. It is a valid assumption therefore that a spacecraft of less than 20 m in size at an altitude over 200 km, where the mean free path exceeds 200 m, is in a free molecular flow regime. As such, the drag force is dependent on the body's shape and size and upon gas—surface interactions that are quantified by the accommodation coefficient, which is a temperature dependent measure of how much energy the flow loses during the interaction. In hyperthermal, free molecular flow regimes, where the speed of the body significantly exceeds the mean molecular speed, the drag coefficient is typically taken between 2.0 and 2.5, depending on the accommodation coefficient, for a wide range of shapes and surface constituents.

It should be noted that at greater altitudes the drag coefficient will increase as the molecular weight of the air decreases and because the assumption of a hyperthermal flow regime breaks down. For altitudes where free molecular flow cannot be assumed, for example, in a reentry scenario when Kn is between about 5 and perhaps as low as

0.1, it is wise to derive a function which defines the drag coefficient as a function of the Knudsen number, as in [16]. Typically the drag coefficient of a spacecraft is determined during the orbit check-out phase in order to aid orbit determination and prediction, but usually only to three significant figures.

4.3.3.4 Effect of Atmospheric Drag and Orbit Lifetime

The principal effect of atmospheric drag is to act as a retarding force against the velocity vector, dissipating orbit energy. If the orbit is initially eccentric, the drag force is maximum at periapsis and acts to circularize the orbit. During the process of circularization it is possible that sufficient orbit energy will be dissipated to cause the orbit to completely decay. Even if this is not the case, once the orbit is circularized dissipation of orbit energy will continue and cause the orbit radius to gradually decay until the body does reenter the atmosphere. General perturbations analysis presented in [16] quantifies the effects of atmospheric drag on orbit size and shape over time, defining four phases of decay due to atmospheric drag and treating them separately

- High eccentricities, $e \gtrsim 0.2$ and $\beta x \gtrsim 30$
- Normal eccentricities, $0.02 \lesssim e \lesssim 0.2$ and $3 \lesssim \beta x \lesssim 30$
- Low eccentricities, $0 < \beta x \lesssim 3$; meaning $0 < e \lesssim 0.02$
- Circular orbits, $e = 0$ and $\beta x = 0$ where

$$x = ae \quad (4.113)$$

and β is defined in Eq. 4.104.

Considering low eccentricity orbits and using subscript 1 to denote initial conditions, with $z = \beta x = ae/H$, as z will be the function of the Bessel function, it can be shown that [16] that

$$\frac{e}{e_1} = \sqrt{1 - \frac{\tau}{\tau_L}} + \mathcal{O}(0.008) \quad (4.114)$$

when $\beta x \lesssim 3$, meaning $0 < e \lesssim 0.02$, $\tau = t - t_1$ and τ_L is taken as

$$\tau_L = \frac{1}{2B'} \left(1 + \frac{H}{2a_1} \left(1 - \frac{9}{20} z_1^2 \right) \right) \quad (4.115)$$

such that

$$B' = \frac{2\pi\delta}{T_1} \rho_{rp,1} x_1 I_1(z_1) e^{-z_1} \quad (4.116)$$

where subscript 1 again denotes initial conditions, δ is defined in Eq. 4.112, x is defined in Eq. 4.113, T_1 is the initial orbit period and $I_1(z_1)$ is a Bessel function of the first kind, of order 1. Equation 4.115 is very nearly the orbit lifetime from $z = z_1$ onwards.

Strictly, Eqs. 4.114 and 4.115 remain valid until $e = 0$, but in practice the periapsis will likely drop below a minimum viable altitude before this happens. It is also prudent to limit any periapsis change to less than about $3H$, because beyond this the scenario should be paused and a new density scale height determined. Failure to do so will overestimate the orbit decay time. As such whilst Eq. 4.115 can approximate the orbit decay time of an initially eccentric orbit due to atmospheric drag when $e \lesssim 0.02$, the change in periapsis should also be considered to determine the validity of using a fixed density scale height. The periapsis distance can be shown [16] to be

$$r_p = r_{p,1} - H \left\{ \left(1 - 3 \frac{H}{a_1} \right) \ln \left(\frac{z_1}{z} \frac{I_1(z_1)}{I_1(z)} \right) - z_1 \left[1 - \frac{2}{5} e_1 \left(1 + \frac{z}{z_1} \right) \right] \left(1 - \frac{z}{z_1} \right) + \alpha \left(\frac{ae^3}{H}, \frac{H}{5a} \right) \right\} \quad (4.117)$$

where

$$z = z_1 \sqrt{1 - \frac{\tau}{\tau_L}} \left\{ 1 + \frac{7H}{4a_1} \left[\frac{z_1^2}{20\tau_L} - \ln \left(\frac{I_1(z_1)}{\sqrt{\left(1 - \frac{\tau}{\tau_L}\right)} I_1\left(z_1 \sqrt{1 - \frac{\tau}{\tau_L}}\right)} \right) \right] \right\} \quad (4.118)$$

The orbit decay time of an initially circular orbit due to atmospheric drag is easily determined from first principles. However, it should be obvious that no orbit is truly circular and hence the above $\beta x \lesssim 3$ conditions give a better approximation of low eccentricity orbit decay under atmospheric drag. Using subscript c to denote initial conditions of the circular orbit, and noting that orbit energy is dissipated by aerodynamic drag and that orbit energy is a function solely of the orbit radius, as seen in Eq. 4.19, it is evident that the rate of energy loss is determined by the dissipated power and the non-dimensional circular orbit lifetime [1] is

$$\frac{\tau_L}{T_c} = \int_0^{\tau_L/T_c} dt = \frac{1}{2\pi} \left(\frac{1}{\rho_c a_c^{3/2}} \right) (F^{-3/2}) \left(\frac{m}{SC_D} \right) \int_{r_L}^{r_c} \frac{\rho_c}{\rho} \frac{dr}{\sqrt{r}} \quad (4.119)$$

where ρ/ρ_c is the exponential atmospheric model from Eq. 4.103 and r_L is the radius at which reentry is taken to occur, typically between 120 and 150 km. From Eq. 4.119 a general approximation can be found by letting $r = r_L + \Delta$, where $1 \gg |\Delta/r_L|$ and $\Delta < 0$, to solve the integral in Eq. 4.119 to obtain the approximate decay time of a circular orbit, where $r_c < 1,000$ km altitude, under atmospheric drag [1] as

$$\tau_L = T_c \left[\frac{1}{4\pi} \left(\frac{2\beta r_c + 1}{\rho_c \beta^2 r_c^3} \right) (F^{-3/2}) \left(\frac{m}{SC_D} \right) (1 - e^{\beta\Delta}) \right]. \quad (4.120)$$

4.3.4 Radiation Pressure

In 1873 the Scottish physicist and mathematician James Clerk Maxwell (1831–1879) predicted the existence of radiation pressure as a consequence of his unified theory of electromagnetic radiation [17]. Independently of Maxwell,

in 1876, the Italian physicist Adolfo Bartoli (1851–1896) also demonstrated the existence of radiation pressure as a consequence of the second law of thermodynamics [18]. Similar to the difficulty of accurately quantifying the surface area of a spacecraft when analyzing atmospheric drag effects, the effect of radiation pressure on a spacecraft can be very difficult to quantify accurately. This difficulty is due to complex time-varying geometries giving uncertainty in illuminated area, self-shadowing, self-illumination due to multiple reflections, variation in time of illuminated material properties due to degradation, and potentially multiple sources of radiation pressure; the main sources of radiation pressure being the Sun and the planets.

Using quantum mechanics, radiation pressure can be visualized as momentum transported by photons impacting and then reflecting off a surface. The term ‘*photon*’ was coined by Gilbert N. Lewis (1875–1946) in a letter to *Nature* magazine in 1926 [19, 20]. From Planck’s Law, a photon of frequency ν will transport the energy given by

$$E = h\nu. \quad (4.121)$$

Using special relativity the total energy of a moving body may be written as

$$E^2 = m_0^2 c^4 + p^2 c^2. \quad (4.122)$$

Since a photon has zero rest mass, its energy may be written as

$$E = pc. \quad (4.123)$$

Using the photon energy defined by Eqs. 4.121 and 4.123, the momentum transported by a single photon is

$$p = \frac{hv}{c}. \quad (4.124)$$

The pressure on a body is found through consideration of the momentum transported by a flux of photons. At distance r from the Sun the energy flux, W , may be written in terms of the solar luminosity, L_S , as

$$W = \frac{L_S}{4\pi r^2}. \quad (4.125)$$

The energy, ΔE , transported across a surface of area A , normal to the incident radiation, in time Δt is given by

$$\Delta E = WA\Delta t \quad (4.126)$$

which from Eq. 4.123 gives the momentum transported as

$$\Delta p = \frac{\Delta E}{c}. \quad (4.127)$$

The pressure on the surface is thus defined as the momentum transported per unit time, per unit area, such that

$$P_{SRP} = \frac{1}{A} \left(\frac{\Delta p}{\Delta t} \right). \quad (4.128)$$

Accordingly, using Eq. 4.125 the pressure exerted on the surface due to momentum transport by photons is

$$P_{SRP} = \frac{W}{c}. \quad (4.129)$$

As such, the radiation pressure exerted on a surface varies as the inverse square of the distance from the radiation source, that is $(1/r^2)$. From Newton's second law, the actual pressure on a perfectly reflecting surface is twice the value given by Eq. 4.129 because momentum is transferred by both incident and by reflected radiation.

4.3.4.1 Solar Radiation Pressure

The $(1/r^2)$ variation of radiation pressure is perhaps most notably observed when considering the effect of solar radiation pressure, SRP, and as such it can be convenient to write Eq. 4.126 as

$$W = W_E \left(\frac{R_E}{r} \right)^2 \quad (4.130)$$

where, $W_E = (L_S/4\pi R_E^2)$ is the energy flux at the Earth's distance from the Sun, $R_E = 1$ au. The luminosity of the Sun is approximately 3.839×10^{26} W (or 3.839×10^{33} erg s⁻¹) [21]. However, it can be slightly higher if solar neutrino radiation is considered as well as

Table 4.3 Solar irradiance at each of the planets, from ASTM E-490

Planet	Solar irradiance (W/m ²)		
	Mean	Perihelion	Aphelion
Mercury	9,116.4	14,447.5	6,271.1
Venus	2,611.0	2,646.4	2,575.7
Earth	1,366.1	1,412.4	1,321.7
Mars	588.6	715.9	491.7
Jupiter	50.5	55.7	45.9
Saturn	15.0	16.8	13.5
Uranus	3.7	4.1	3.4
Neptune	1.5	1.5	1.5

electromagnetic radiation. Additionally, the Sun is a weakly variable star. The principal fluctuation is due to the 11-year solar cycle, giving a periodic variation in luminosity of about $\pm 0.1\%$. From the solar luminosity the energy flux at the Earth's mean distance from the Sun, or solar irradiance, can be determined as approximately 1,366 W/m²; which corresponds to the integrated power from ASTM E490-00a(2006) and ISO-21348 as discussed in Chap. 3. However, the actual 'best' solar irradiance at the Earth's distance for use in engineered systems remains subject to engineering judgment, as some values quoted are as high as 1,377 W/m². It is therefore often best to select a solar irradiance that gives a conservative design estimate. For example, in Chap. 10 the conservative value of 1,353 W/m² is discussed for use in power system design.

The acceleration due to SRP can be determined using a reflectivity coefficient, C_R , such that

$$\mathbf{F}_{SRP} = - \frac{P_{SRP} C_R A_{SRP}}{m} \frac{\mathbf{r}_{SRP}}{|\mathbf{r}_{SRP}|} \quad (4.131)$$

where A_{SRP} is the surface area exposed to the solar radiation and \mathbf{r}_{SRP} is a vector from the spacecraft to the Sun. As such the acceleration due to SRP is always directed away from the Sun. The reflectivity coefficient, C_R , takes a value between zero and two and characterizes how the incoming radiation is reflected. If C_R is zero the body is transparent to incoming radiation, if C_R is one all incoming radiation is absorbed, that is, it is a black-body with zero reflection, and if C_R is two the body is a perfect reflector. A perfectly reflective surface facing the Sun at Earth's mean distance from the Sun will experience a pressure of just over 9 μ Pa. The value of C_R is clearly critical in determining the perturbing acceleration magnitude due to SRP. However, due to the difficulties outlined previously it is almost impossible to predict what value of C_R a spacecraft will have. Instead, just as with the spacecraft drag coefficient, C_R is typically determined from in-flight data. It should also be noted that for solutions that are slightly more detailed, the SRP

Table 4.4 Geometric and bond albedo at each of the planets and the Moon

Planet	Geometric albedo	Bond albedo
Mercury	0.14	0.07
Venus	0.67	0.90
Earth	0.37	0.31
Moon	0.12	0.11
Mars	0.17	0.25
Jupiter	0.52	0.34
Saturn	0.47	0.34
Uranus	0.51	0.30
Neptune	0.41	0.29

acceleration can be split into three components, absorbed radiation, and specular and diffusely reflected radiation.

Finally, it should be noted that as the orbit of the Earth about the Sun is slightly elliptical, solar irradiance at the Earth, and hence radiation pressure, varies by approximately 3.5 % over the year. The mean, maximum and minimum solar irradiance at each planet is found in Table 4.3.

The effect of SRP can be significant in certain orbit regimes and negligible in others, it can also often average to zero over the orbit period. Depending on the orbit regime, the orbit size, shape and orientation, and spacecraft parameters, the effect of SRP can be exploited to enhance a spacecraft's performance. A well-known example of this is the concept of solar sailing, where a large lightweight reflective surface provides a propulsive thrust for the spacecraft to perform orbit maneuvers [22, 23]. A similar concept is widely used for attitude control of spacecraft in both Earth orbit and for inner solar system missions, where due to the $(1/r^2)$ variation in radiation pressure, significant forces can be generated. Both the Mariner-10 and MESSENGER spacecraft used SRP by design to offset propellant requirements, while the Hayabusa spacecraft was able to use SRP to recover from a partial failure of its attitude control systems. At Earth it is common to use so-called 'trim tabs' to aid attitude control of large spacecraft in geostationary orbit, thereby reducing propellant requirements and extending mission lifetime. Trim tabs are typically mounted onto the solar arrays in order to maximize the torque delivered.

4.3.4.2 Planetary Albedo and Infrared Radiation

Of the solar radiation that impinges on a body, a certain fraction is reflected from that body and the ratio of reflected to impinging radiation is termed the body's albedo. In astronomy two measures of reference are commonly used, geometric and Bond albedo. The geometric albedo of an

astronomical body is a measure of its brightness when illuminated from a phase angle behind the observer, whilst the Bond albedo is a measure of the total proportion of electromagnetic radiation reflected. The geometric and Bond albedo of each planet, and the Moon, is given in Table 4.4. The remaining solar radiation is absorbed and re-emitted at a later time as infrared radiation; from Earth this amounts to about 237 W/m^2 . Due to atmospheric absorption at specific wavelengths, it is common to split the effects of planetary albedo and infrared radiation into specific wavelengths. At Earth, the effects of albedo can typically be neglected for all but high-accuracy calculations, but at other bodies, especially those without atmospheres, the effect can be significant. The highest known albedo in the solar system is found at the Saturnian moon Enceladus, with an albedo of over 0.99 because it is essentially fresh snow.

4.3.5 Minor Forces

A myriad of minor forces act on a spacecraft, but typically these can be neglected.

4.3.5.1 Tides

Tides are a result of gravitational effects causing a distortion in a body's mass distribution. The most apparent tidal motion at Earth is that of the oceans, but the most astrodynamically significant tidal effect is termed 'Solid-Earth tides', which are deformations of the Earth's shape due, principally, to the gravitational attraction of the Moon and the Sun. It should also be noted that forces in the Earth's interior can contribute to Solid-Earth tides, as can the centrifugal effect of the Earth's rotation. The effect of tides is typically determined for Earth orbiting spacecraft through analysis of flight data, as it is not directly observable.

4.3.5.2 Solar wind

The solar wind is a stream of charged particles ejected from the Sun and is distinct from solar radiation pressure. The solar wind exerts a pressure approximately four orders of magnitude less than direct solar radiation pressure.

4.3.5.3 General Relativity Effects

Although not strictly a perturbing force, General Relativity (GR) dictates that light travels in a curved path in space due to massive particles. As such, when measuring the position of distant galaxies or quasars, GR effects must be considered, as the light will have been 'deflected' by massive objects it has passed *en route*. For spacecraft, this angular deflection is typically negligible. However, in certain scenarios it can cause an apparent spacecraft deviation from a propagated trajectory.

4.4 Orbit Classifications

The objectives of a space mission typically drive the mission design towards the use of certain orbits. In addition to this, consideration of orbit parameters and perturbations leads to the definition of special orbits, with specific desirable characteristic. As such, it is convenient to define different orbit classifications; here these classifications are given specifically for Earth, but it should be apparent that equivalent orbits may exist for any central body.

4.4.1 Low Earth Orbit

Generally classified as extending from the Von Kármán ellipsoid to below the peak radiation levels of the inner van Allen belt, the Low Earth Orbit (LEO), regime is specifically defined by inter-agency agreement to be the altitude range 160–2,000 km. Due to the partial-vacuum nature of this region, all orbits in LEO experience the effects of the upper atmosphere, including atmospheric drag, but as discussed in the preceding section beyond approximately 1,000 km altitude the effects can be marginal. Additionally, due to atmospheric drag effects the altitude is typically greater than 300 km.

Whilst atmospheric drag is the most apparent consequence of the upper atmosphere, chemical interactions between the atmosphere and the materials on the spacecraft should also be considered; particularly due to atomic oxygen. Indeed, oxidization due to atomic oxygen requires that spacecraft in LEO be covered in a non-oxidizing material. Furthermore, due to protection by the atmosphere, and its location below the peak radiation levels of the inner van Allen belt, the LEO regime exposes spacecraft to relatively low levels of radiation. However, the short orbit period potentially exposes the spacecraft to frequent and lengthy passages through the Earth's shadow, resulting in thermal shocks and cycling.

By agreement of the Inter-Agency Space Debris Coordination Committee, LEO is one of two designated space debris protected orbit regimes; the other is the geosynchronous region. Spacecraft are generally required to exit protected orbit regimes within 25 years of the end of life.

4.4.2 Medium Earth Orbit

Extending from LEO, the Medium Earth Orbit (MEO) regime extends to below the geosynchronous region. It is noteworthy that the MEO includes near-circular orbits of 12 h period, an example ground track of which is shown in Fig. 4.9. One of the most common applications within the

MEO regime is constellations of spacecraft providing a Global Navigation Satellite System (GNSS) services. Typically, spacecraft in MEO are located beyond the inner van Allen belt but are more exposed to the outer van Allen belt than geosynchronous spacecraft, and thus suffer a higher electron flux.

4.4.3 Geosynchronous Orbit

A geosynchronous orbit is an Earth orbit with period equal to the Earth's sidereal rotation period, with restrictions on orbit inclination and eccentricity. By agreement of the Inter-Agency Space Debris Coordination Committee space debris in the geosynchronous region is restricted to a maximum of 15° geodetic latitude and to within 200 km of the geostationary altitude.

4.4.3.1 Geostationary Orbit

A geostationary orbit (GEO) is idealized as a circular geosynchronous orbit with zero inclination; the altitude of this orbit is defined as 35,786 km using the equatorial Earth radius. An object in GEO appears stationary to an observer on the ground. Typically, spacecraft considered to be in GEO do not have a perfect GEO orbit, but will have a slight inclination and/or eccentricity, causing the spacecraft to depict a lemniscate curve as shown (in exaggeration for a GEO) in Fig. 4.9 for a 30° inclined geosynchronous orbit. Geostationary orbits are widely used for communications and Earth observation applications. However, as the observed geodetic latitude is increased the observation zenith angle increases such that Earth observation applications become restricted beyond approximately 55° latitude, and communication applications beyond approximately 70° latitude. The orbit environment in GEO is relatively benign, but it is still within the outer van Allen belt and hence some radiation effects must be considered.

4.4.3.2 Geostationary Transfer Orbit

A geostationary transfer orbit (GTO) is an orbit used to transfer spacecraft into the geosynchronous region. A GTO typically has its periapsis within the LEO region and its apoapsis near GEO. Most spacecraft designed to operate in GEO are inserted into a GTO by the launch vehicle and then progress to GEO using an on-board propulsion system. The orbit environment in GTO is very harsh due to the short orbit period that exposes the spacecraft to much more frequent passages through the Earth's shadow than during its operational life, resulting in thermal shocks and cycling, as well as passages through both the inner and outer van Allen belts twice per orbit.

4.4.4 High Earth Orbit

A High Earth Orbit is an orbit with its apoapsis altitude more than 200 km beyond that of a geostationary orbit.

4.4.4.1 Highly Elliptical Orbit

A Highly Elliptical Orbit (HEO) is a subset of the High Earth Orbit classification in which the orbit eccentricity places the periapsis within LEO. Such high eccentricities, through Kepler's second law, have long dwell times around apoapsis, where they appear almost stationary to a ground-based observer. An example of a service delivered from HEO is the Sirius Satellite Radio service in North America. Radiosat-1, -2, and -3 were placed into 24 h HEOs, and Radiosat-5 into a conventional GEO. Radiosat-4 was built as a ground spare for Radiosat 1-3 and held in storage until 2012, when it was transferred to the Smithsonian Institution's National Air and Space Museum in Washington, D.C. As with GTOs, HEOs can experience quite a severe space environment.

4.4.5 Sun-Synchronous Orbits

Through careful consideration of the orbit perturbation force due to a non-spherical central body, a secular variation of the ascending node angle of a near-polar orbit can be induced without expulsion of propellant. Consequently, the orbit perturbations can be used to maintain the orbit plane in a fixed orientation with respect to the Sun-line throughout the full year of the primary body; such orbits are termed Sun-synchronous orbits. Sun-synchronous orbits about the Earth are typically near-circular LEOs, with an altitude of less than 1,500 km. It is normal to design a LEO such that the orbit period is synchronised with the rotation of the Earth's surface over a given interval, such that a repeating ground track is established. A repeating ground track, together with the near-constant illumination conditions of the ground track when observed from a Sun-synchronous orbit, enables repeat observations of a target over an extended time under similar illumination conditions; for this reason, Sun-synchronous orbits are used extensively by Earth observation platforms.

Recalling that the dominant perturbation due to the Earth's shape is J_2 , which is three orders of magnitude larger than J_3 , Eq. 4.97 can, for a body possessing axial symmetry, be written as

$$U(r, \beta) = \frac{\mu}{r} \left[1 - \sum_{n=0}^{\infty} J_n \left(\frac{R_{\oplus}}{r} \right)^n P_n \sin \beta \right] \quad (4.132)$$

where R_{\oplus} is the radius of the Earth. It is thereafter found that

$$\begin{aligned} U(r, \beta) = & \frac{\mu}{r} \left[1 - J_2 \frac{1}{2} \left(\frac{R_{\oplus}}{r} \right)^2 (3 \sin^2 \beta - 1) \right. \\ & - J_3 \frac{1}{2} \left(\frac{R_{\oplus}}{r} \right)^3 (5 \sin^3 \beta - 3 \sin \beta) \\ & \left. - J_4 \frac{1}{8} \left(\frac{R_{\oplus}}{r} \right)^4 (3 - 30 \sin^2 \beta + 35 \sin^4 \beta) - \dots \right] \end{aligned} \quad (4.133)$$

Using spherical triangle laws and considering only the first-order terms, Eq. 4.133 reduces to

$$\begin{aligned} U(r, \beta) = & U_o + U_p \\ = & \frac{\mu}{r} - J_2 \frac{\mu R_{\oplus}^2}{2r^3} (3 \sin^2 i \sin^2(\theta + \omega) - 1). \end{aligned} \quad (4.134)$$

A Sun-synchronous orbit requires that the rate of change of the ascending node match the mean rate of rotation of the Sun within an Earth-centred inertial reference frame. The ascending node angle is described in the Gaussian form in Eq. 4.42, and by modifying this such that the position-fixing element is the true anomaly, the rate of change of the ascending node angle becomes

$$\frac{d\Omega}{d\theta} = \frac{r^3}{\mu p} [R \quad T \quad N] \begin{bmatrix} 0 \\ 0 \\ \left(\frac{\sin(\theta + \omega)}{\sin i} \right) \end{bmatrix}. \quad (4.135)$$

Thus, U_p within Eq. 4.134 is required in terms of the satellite-based Gaussian coordinate system, or simply RTN; see Sect. 4.1.5. This is obtained by differentiation of the potential with respect to the spacecraft centred RTN coordinate system. The disturbing force components due to J_2 are thus

$$R_{J_2} = \frac{3}{2} J_2 \frac{\mu R_{\oplus}^2}{r^4} (3 \sin^2 i \sin^2(\theta + \omega) - 1) \quad (4.136)$$

$$T_{J_2} = -\frac{3}{2} J_2 \frac{\mu R_{\oplus}^2}{r^4} \sin^2 i \sin^2(\theta + \omega) \quad (4.137)$$

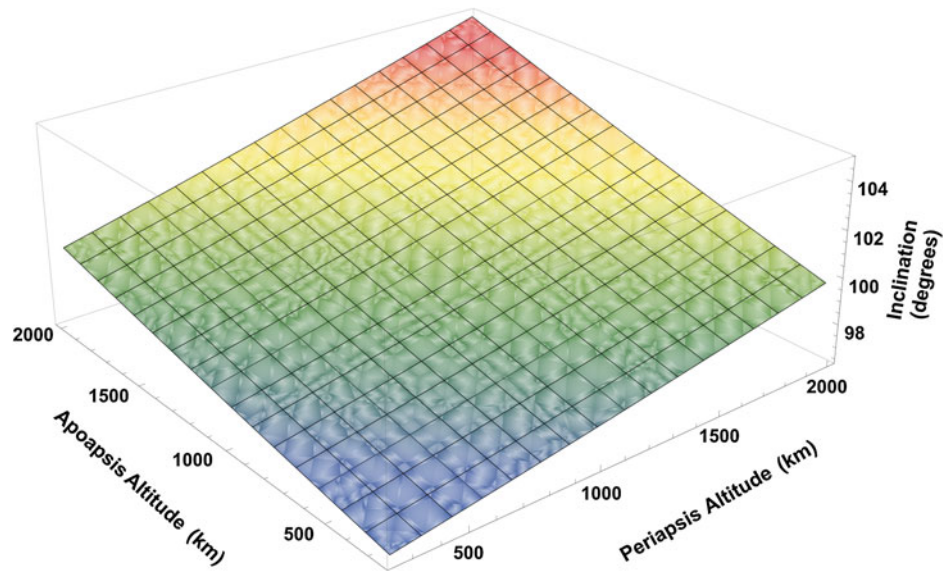
$$N_{J_2} = -\frac{3}{2} J_2 \frac{\mu R_{\oplus}^2}{r^4} \sin 2i \sin(\theta + \omega). \quad (4.138)$$

By combining Eqs. 4.135 and 4.138, and assuming that the change in other orbit elements is small over the integral, the Sun-synchronous orbit can be found as

$$i = \cos^{-1} \left[-\frac{2 \Delta\Omega a^{7/2} (1 - e^2)^2}{3 J_2 R_{\oplus}^2 \sqrt{\mu}} \right] \quad (4.139)$$

where $\Delta\Omega$ is the mean rotation rate of the Sun within an Earth-centred inertial reference frame per second. Eq. 4.139 is solved in Fig. 4.13 for a range of periapsis and apoapsis altitudes.

Fig. 4.13 Surface of Sun-synchronous orbits within LEO.
Image Malcolm Macdonald



4.4.6 Critical Inclination Orbits

Through further careful consideration of the orbit perturbation force due to a non-spherical central body, it is noted that the induced secular variation of the argument of periapsis is inclination dependent. The argument of periapsis angle is described in the Gaussian form in Eq. 4.43, and by modifying this such that the position-fixing element is the true anomaly, the rate of change of the ascending node angle becomes

$$\frac{d\omega}{d\theta} = \frac{r^2}{\mu e} [R \quad T \quad N] \begin{bmatrix} -\cos \theta \\ \left(1 + \frac{r}{p}\right) \sin \theta \\ \frac{re}{p \tan i} \sin(\theta + \omega) \end{bmatrix}. \quad (4.140)$$

By combining Eqs. 4.136–4.138 with Eq. 4.140, and assuming that the change in other orbit elements is small over the integral, the expression for the change in argument of periapsis is found to be

$$(\Delta\omega)_0^{2\pi} = \frac{3J_2 r^2 (3 + 5 \cos(2i))}{4a^2 (-1 + e^2)^2}. \quad (4.141)$$

Seeking zero secular variation of the argument of periapsis, Eq. 4.141 is solved equal to zero; which occurs when $(3 + 5 \cos(2i)) = 0$. Consequently, the critical inclination at Earth is determined as $90 \pm 26.6^\circ$. Thus to the order of J_2 , all Earth orbits inclined at these values show no rotation of the apsidal line, irrespective of the values of semi-major axis and eccentricity.

4.4.6.1 Molniya Orbit

Named after the series of Soviet satellites which used the orbit, a Molniya (meaning ‘lightning’) orbit is a type of HEO inclined at the critical inclination, with an orbit period of one half of a sidereal day. An example ground track of a Molniya orbit is shown in Fig. 4.9. By coupling the apogee dwell features of a HEO with critical inclination, this orbit overcomes the difficulties of communicating with high latitude regions from a GEO. As with most HEOs, Molniya orbits can experience quite a severe space environment.

4.4.6.2 Tundra Orbit

This is a type of geosynchronous HEO inclined at the critical inclination. As with other geosynchronous orbits, the ground track of a tundra orbit is a lemniscate curve.

4.5 Trajectory Maneuvering

The requirement to maneuver from one orbit to another is, by definition, a critical study within astrodynamics because it is through the addition of “*artificially induced forces*” that astrodynamics is distinct from its parent sciences into the field of engineering.

The problem of maneuvering from an initial vector-defined position to a target vector-defined position in a given time is known as Lambert’s Problem and is applied, for example, to target a flyby or rendezvous with a target body. However, for preliminary trajectory maneuvering analysis it

Fig. 4.14 Tangential burn at an apsis. *Image Malcolm Macdonald*

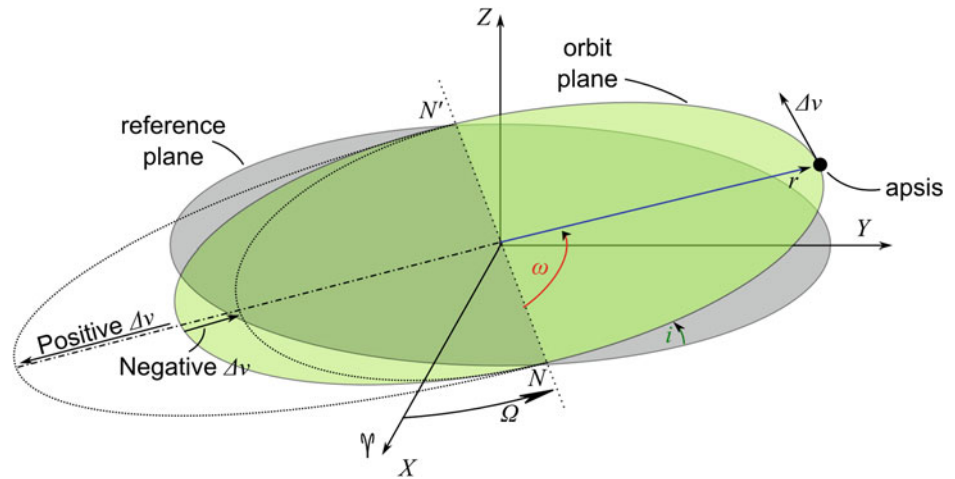
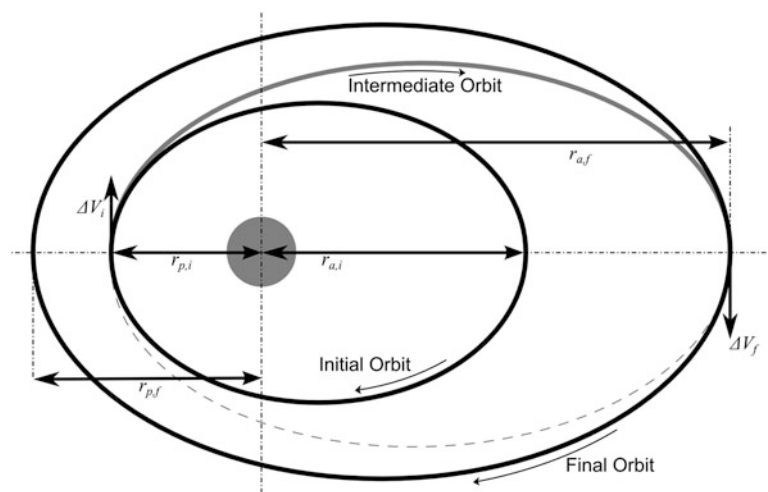


Fig. 4.15 Hohmann transfer to a larger semi-major axis orbit. *Image Malcolm Macdonald*



is often sufficient to consider only the required velocity change, negating the requirement to solve Lambert's Problem. Thus Lambert's Problem is not addressed further herein.

4.5.1 Coplanar Maneuvers

It is self-evident that coplanar maneuvers can alter the orbit semi-major axis, eccentricity and argument of periapsis; this can also be seen mathematically in Eqs. 4.39–4.43. Coplanar burns are either tangential (to the orbit) or non-tangential. From an efficiency perspective tangential burns are preferred, but non-tangential burns can shorten the transfer time between two orbits at the expense of requiring a larger burn and hence more fuel.

Consider a spacecraft at an apsis of an initial orbit. If a propulsive maneuver is made to alter the spacecraft's velocity magnitude the effect is to move the opposite apsis, as shown in Fig. 4.14. This can be understood by inserting Eqs. 4.5 and 4.6 into Eq. 4.17 to gain the velocity at each apsis. By increasing the velocity at periapsis the velocity at

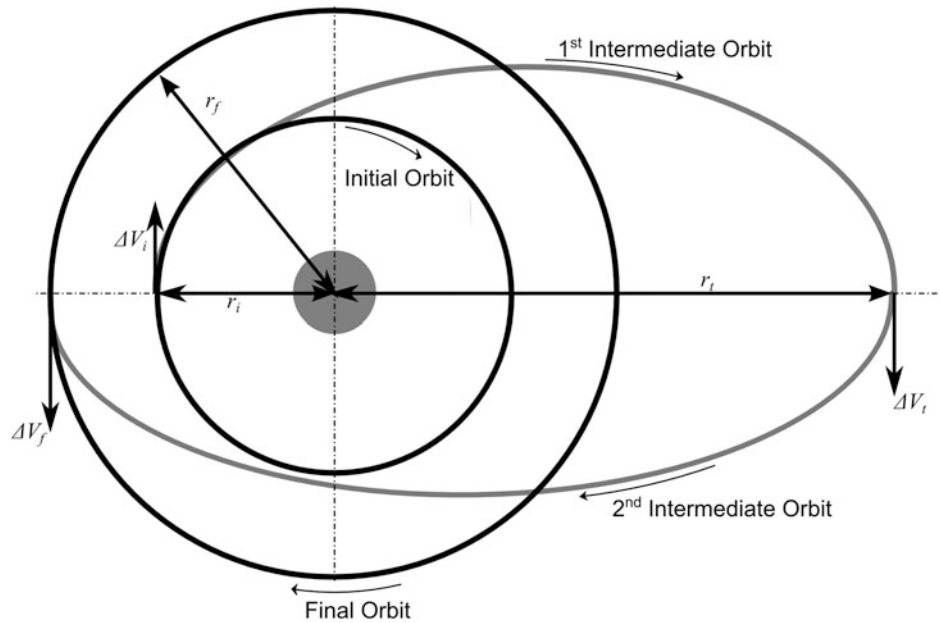
apoapsis is reduced, and vice versa. Thus, a spacecraft can move between two orbits that are tangential at an apsis by a single impulsive maneuver at that apsis. The size of the velocity change maneuver that is required is defined by the target change of radius of the opposite apsis. An example of such a maneuver is that of a spacecraft inserted into a GTO by the launch vehicle, which then progresses to GEO by using an on-board propulsion system to alter its velocity at the GTO apoapsis.

4.5.1.1 Hohmann Transfer

The simplest and most common approximation of a trajectory maneuver between two coplanar and coaxial orbits, which are not tangential, is the Hohmann transfer, as this typically provides the minimum required change in velocity. A Hohmann transfer joins two coplanar and coaxial orbits, which are not tangential, through an intermediate orbit that is tangential to both the initial and final orbit.

A Hohmann transfer maneuver consists of two impulsive velocity changes, the first to transfer onto the intermediate orbit, and the second to transfer from this to the final orbit,

Fig. 4.16 Bi-Elliptic transfer between two circular orbits.
Image Malcolm Macdonald



as illustrated in Fig. 4.15. Note that neither the initial nor the final orbit need be circular. As before, the size of each maneuver that is the required change in velocity is defined by the target change of radius of the opposite apsis. As such, in Fig. 4.15 the initial maneuver, ΔV_i , increases the apoapsis radius to the target value, whilst the second maneuver, ΔV_f , increases the periapsis radius to the target value.

Using Eq. 4.17 the total velocity change can be found as

$$\Delta V = |V_i| + |V_f| \quad (4.142)$$

where

$$V_i = V_{p,t} - V_{p,i} = \sqrt{\mu \left(\frac{2}{r_{p,t}} - \frac{1}{a_t} \right)} - \sqrt{\mu \left(\frac{2}{r_{p,i}} - \frac{1}{a_i} \right)} \quad (4.143)$$

and

$$V_f = V_{a,f} - V_{a,t} = \sqrt{\mu \left(\frac{2}{r_{a,f}} - \frac{1}{a_f} \right)} - \sqrt{\mu \left(\frac{2}{r_{a,t}} - \frac{1}{a_t} \right)} \quad (4.144)$$

with subscript i denoting the initial orbit, f the final orbit, and t the transfer or intermediate orbit, such that $a_t = (r_{p,i} + r_{a,f})/2$. Finally, note that the orbit maneuver duration is simply half of the period of the intermediate orbit, which from Eq. 4.14 is $\tau_{transfer} = \pi \sqrt{a^3/\mu}$.

4.5.1.2 Bi-Elliptic Transfers

When the required change in orbit radius is large, the Hohmann transfer is found to be sub-optimal [4, 24]. In this case, a third orbit maneuver is introduced as shown in

Fig. 4.16, and the result is termed a bi-elliptic transfer. In a bi-elliptic transfer the initial maneuver occurs at the same location as in a Hohmann transfer, but the magnitude of the velocity change is increased such that the apoapsis of the intermediate orbit is larger than that of the target, or final orbit. A second maneuver is performed at the apoapsis of the intermediate orbit to create a second intermediate orbit with a periapsis matching that of the target value. The periapsis of the second intermediate orbit is thus tangential with the target orbit, at which point a third maneuver is performed to acquire the final orbit. The orbit maneuver duration is the sum of half of the orbit period of each of the intermediate orbits.

A Hohmann transfer between two circular, co-planar orbits is found to be optimal (minimum change in total velocity, and hence minimum fuel mass) when the ratio of initial to final orbit radius is <11.94 [24]. Meanwhile, a bi-elliptic transfer between two circular, co-planar orbits can be shown optimal when the ratio of initial to final orbit radius is >15.58 [24]. When the ratio of initial to final orbit radius is between 11.94 and 15.58, further analysis is required to determine the intermediate radius, r_t , at which the bi-elliptic transfer becomes optimal. Note, at orbit radius ratio 11.94, $r_t \rightarrow \infty$ and the intermediate orbit is a parabola.

4.5.2 Plane Change

The orbit plane is perpendicular to the angular momentum vector, \mathbf{h} . As such, a pure plane change maneuver is equivalent to a change in direction of the angular momentum vector through the application of a torque to the orbit

plane. From a simple vector analysis, the change in required velocity, assuming $v_1 = v_2$, is

$$\Delta V = 2V \sin\left(\frac{\zeta}{2}\right) \quad (4.145)$$

where ζ is the plane change angle; for example, a change in orbit inclination.

4.5.3 Ideal Rocket Equation

The required propellant mass for a given change in velocity can be determined through conservation of linear momentum and by defining the Specific Impulse, I_{sp} , of a propulsion system as the ratio of thrust to Earth-surface weight flow rate, $\dot{m}g$, where \dot{m} is the propellant mass flow rate. That is, $I_{sp} = \mathbf{F}/\dot{m}g = v_e/g$, where \mathbf{F} is propulsive thrust and v_e is the exit velocity of the propulsion system exhaust gas; units of I_{sp} are seconds. By definition, I_{sp} is a measure of the energy content of the propellant and how efficiently it is converted into thrust.

The ideal rocket equation gives the available change in velocity by a spacecraft as

$$\Delta V = gI_{sp} \ln\left(\frac{m_0}{m_0 - m_p}\right) = gI_{sp} \ln\left(\frac{m_0}{m_f}\right) \quad (4.146)$$

where m_0 is the initial mass of the spacecraft and m_p is the propellant mass, such that the final mass is $m_f = (m_0 - m_p)$. Note that this is an ideal case with no losses and as such represents the ideal limiting case. It can be useful in preliminary studies to determine the fuel mass fraction, m_p/m_0 , required to deliver a given change in velocity. As the fuel mass fraction tends to one, the allowable spacecraft dry mass (spacecraft mass without propellant) tends to zero. The fuel mass fraction is

$$\frac{m_p}{m_0} = 1 - e^{-(\Delta V/I_{sp}g)}. \quad (4.147)$$

4.5.4 Finite Burn Losses

The above assumption that the change in velocity is delivered by a single impulsive maneuver breaks down when considering large maneuvers where, in practice, the spacecraft thruster(s) may fire for several tens of minutes or more in order to deliver the required thrust and hence change in velocity. Consequently, the thrust is delivered over an arc of the trajectory, typically centered about an apsis. This spatial distribution of the thrust results in an inefficiency in the use

of propellant and is termed ‘finite burn’ losses. To counteract finite burn losses the required change in velocity is often delivered over a number of orbits in order to minimize the arc over which thrust is delivered. However, such a strategy can have a significant effect on the time taken to complete the maneuver.

4.5.5 Continuous Thrust

Through the application of low-thrust propulsion systems, such as solar electric propulsion, it is possible to apply continuous thrust to a spacecraft. Such trajectories differ significantly from those of high-thrust spacecraft, with only small changes in the element set per orbit.

Adopting the assumption of continuous thrusting by a spacecraft in an initially circular Keplerian orbit enables analytical solutions to be gained for some general problems. For example, when the thrust vector is directed either radially or tangentially general analytical results can be found. This principle was apparently first used by Tsien [25] who considered the use of both radially and tangentially constant acceleration in terms of “take off” from an orbit, showing that tangentially constant acceleration is much more efficient than radial thrusting in gaining orbit energy because the required mass ratio is much reduced. Analytical solutions were also developed for transfers between inclined circular orbits by Edelbaum [26], the results of which allow for estimation of the velocity increment (ΔV) and the transfer time for missions with continuous low-level thrusting. This work has since been extended to remove the need for numerical integration of differential equations and to accurately accommodate the effects of periods of zero thrust that may be encountered due to power unavailability during, for example, passage through the Earth’s shadow when using solar electric propulsion [27–29].

4.5.5.1 Low-Thrust Orbit Raising

Consider a spacecraft in a quasi-circular orbit, thrusting continuously along the velocity vector (or against it for orbit lowering). Noting that the work done by the propulsion system will increase the orbit energy, the trajectory becomes an outward quasi-circular spiral due to the continuous thrust. Assuming a constant acceleration, and hence a short transfer with high specific impulse giving a (near-)constant mass spacecraft, and recalling the orbit energy equation, Eq. 4.19, the effective change in velocity from initial to final orbit can be shown to equal the change in circular orbit velocity. Moreover, the transfer time can be approximated as $\Delta t = \Delta V/a_{prop}$, where $a_{prop} = \mathbf{F}_{prop}/m$ is the acceleration due to the continuous thrust, \mathbf{F}_{prop} .

4.5.6 Gravity Assist

The trajectory of a spacecraft can be altered using the relative motion of a large celestial body to deflect the velocity vector of the spacecraft, hence in effect providing a non-tangential maneuver without the use of propellant. The first gravity assist maneuver was performed by the Mariner-10 probe at Venus on February 5, 1974. The Cassini-Huygens mission used Venus twice, then Earth and finally Jupiter *en route* to Saturn, giving a transfer duration of 6.7 years. A Hohmann transfer would have required only 6 years, but the total propulsive change in velocity required to be delivered by the spacecraft by using gravity assist maneuvers was reduced from about 16 km/s for a Hohmann transfer to only 2 km/s.

Assuming a symmetric hyperbolic pass, conservation of energy requires the spacecraft hyperbolic excess velocity, V_∞ , be the same on both the arrival and departure asymptotes. The spacecraft hyperbolic excess velocity, V_∞ , that is the velocity relative to the planet, can be found from the *vis-viva equation*, Eq. 4.17, with $r \rightarrow \infty$

$$V_\infty^2 = -\frac{\mu}{a} = 2E_T. \quad (4.148)$$

Note however that the vector velocity \mathbf{V}_∞ is altered by the encounter due to a change in its direction.

The true anomaly of the arrival asymptote can be found from the equation of a polar equation of a conic section, Eq. 4.4, as

$$\theta_a = \cos^{-1}\left(\frac{-1}{e}\right). \quad (4.149)$$

The change in direction of \mathbf{V}_∞ from θ_a is denoted by Ψ , as shown in Fig. 4.17. The departure asymptote true anomaly is thus

$$\theta_d = (\pi - \theta_a + \Psi). \quad (4.150)$$

Hence, if the motion were unperturbed during the flyby the departure asymptote would have true anomaly $(\pi - \theta_a)$. Using Eq. 4.148, the change in direction of \mathbf{V}_∞ can be written as

$$\frac{\Psi}{2} = \theta_a - \frac{\pi}{2} = \sin^{-1}\left(\frac{-1}{e}\right). \quad (4.151)$$

Using Eq. 4.6 and 4.147, the eccentricity can be written as

$$e = 1 + \frac{V_\infty^2 r_p}{\mu}. \quad (4.152)$$

The resultant change in velocity can be written as

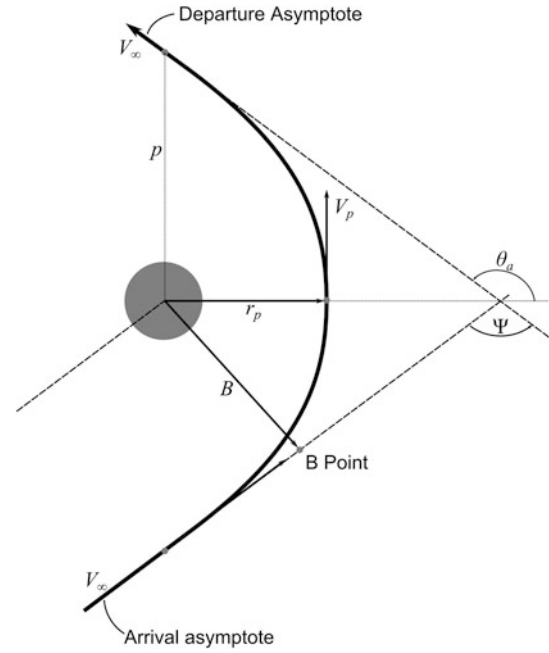


Fig. 4.17 Gravity assist B-plane and B-point. *Image* Malcolm Macdonald

$$\Delta V = 2V_\infty \sin\left(\frac{\Psi}{2}\right) = \frac{2V_\infty}{e}. \quad (4.153)$$

The maximum change in velocity is found by maximizing the deflection angle, Ψ , which is maximized by minimizing the flyby distance.

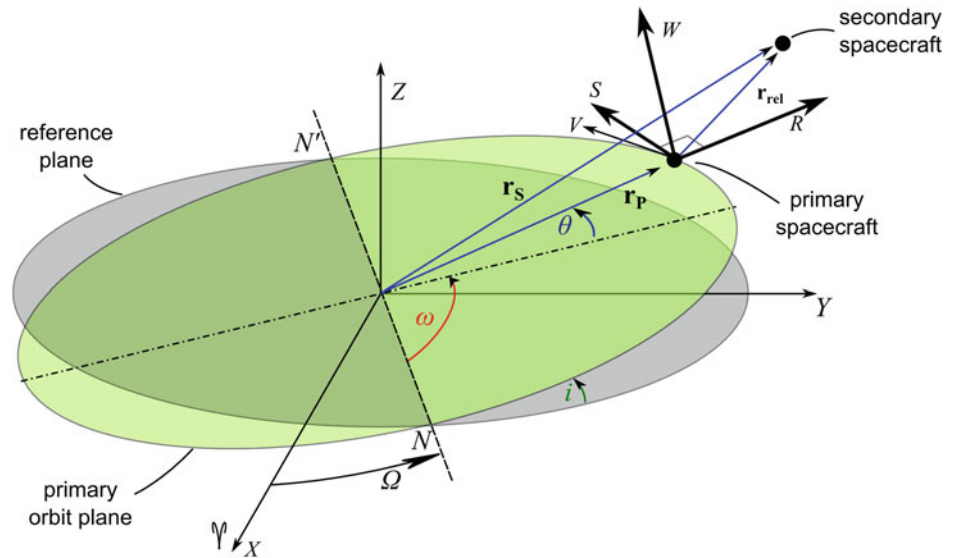
A gravity assist maneuver is targeted using the B-plane, defined as the plane perpendicular to the hyperbolic excess velocity vector and the B-point, the point in the B-plane that the spacecraft would pass through if the gravity of the body being flown past were neglected; as shown in Fig. 4.17. From conservation of angular momentum, $\mathbf{r}_p \mathbf{V}_p = B \mathbf{V}_\infty$ as all vectors are perpendicular, and noting that the velocity on a hyperbolic orbit is $V^2 = [(2\mu/r) + V_\infty^2]$ the B-point can be related to the flyby periapsis as

$$B = r_p \sqrt{1 + \left(\frac{2\mu}{V_\infty^2 r_p}\right)} \quad (4.154)$$

where μ is the gravitational parameter of the body providing the assist. Thus, for a given B-point, the resultant periapsis radius can be determined and the risk of collision with the body assessed.

Due to the limitation on minimum approach, and the resultant limitation on velocity change, an additional change in velocity can be gained by firing thrusters at the closest approach point, or by using aerodynamic surfaces to generate lift and drag; this latter option is termed an

Fig. 4.18 Relative motion geometry, not to scale. *Image* Malcolm Macdonald



aeroassist. When considering gravity assist trajectories it is often convenient to use the method of ‘patched conics’ to divide the trajectory into manageable segments. As the name suggests, the patched conics method simply patches together multiple conic sections based on logical phases, for example the entry or exit of a sphere of influence as shown in Fig. 4.17.

4.5.7 Aerobraking

The concept behind aerobraking is to use atmospheric drag to aid the process of capturing a spacecraft about a target body. Using a low periapsis, the apoapsis is reduced to the desired point and then the spacecraft propulsion system is used to increase periapsis and thus stop orbit decay due to atmospheric drag. Aerobraking was first performed by the Hiten spacecraft using the Earth’s atmosphere. The first aerobraking maneuver away from the Earth was performed by the Magellan spacecraft during its mission extension phase at Venus. The first spacecraft to use aerobraking as the primary form of capture and orbit lowering was Mars Global Surveyor, which used its solar panels to control the aerobraking process. By using aerobraking the Mars Global Surveyor spacecraft propellant requirements were reduced by over 220 kg. Note however that some saved propellant mass is required for the aerobraking assembly and thermal protection system.

4.5.8 Formation Flying and Rendezvous

When considering the close relative motion of spacecraft, such as close formation flying or the terminal phase of rendezvous maneuvers, a reference frame centered on the

primary body of the system, i.e. Sun or Earth, is typically inappropriate as the differential acceleration on each spacecraft is small and fine relative motion may be masked by orbital motion. In this case, it is standard to directly describe the relative motion of one spacecraft in a non-inertial reference frame centered on the other spacecraft; this has the additional advantage for rendezvous scenarios, of the guidance being described relative to the target spacecraft.

The equations of relative motion are given in the satellite-based Gaussian, or RSW coordinate system introduced previously, where as shown in Fig. 4.18 the origin is located at the target, or primary spacecraft. Denoting a spacecraft as the target or primary is of arbitrary importance and is typically only a matter of context. The equations of motion of the primary spacecraft are, from Eq. 4.3 and using two-body motion, given as

$$\ddot{\mathbf{r}}_P = -\frac{\mu}{r_P^3} \mathbf{r}_P. \quad (4.155)$$

The secondary spacecraft can be described similarly, however it will be subject to some differential acceleration, $\mathbf{F} = [f_x \ f_y \ f_z]$, and is thus akin to Eq. 4.37

$$\ddot{\mathbf{r}}_S = -\frac{\mu}{r_S^3} \mathbf{r}_S + \mathbf{F}. \quad (4.156)$$

Assuming that the relative vector, \mathbf{r}_{rel} , from the primary spacecraft to the secondary spacecraft is small, less than approximately 10 % of the orbit radius, and noting that it is given as $\mathbf{r}_{rel} = (\mathbf{r}_S - \mathbf{r}_P)$, the relative equation of motion of the secondary spacecraft can be derived through the use of vector identities after differencing Eqs. 4.155 and 4.156. Thereafter, the linearized form of the equations of motion depends on the applied simplifying assumptions. The most common simplifying assumption made is that both the

primary and secondary spacecraft are in quasi-circular orbits, with similar semi-major axis, inclination and ascending node. George William Hill (1838–1914) first derived the relative equations of motion using these assumptions in 1877 within the context of celestial mechanics [30]; they were subsequently re-derived by W. H. Clohessy and R. S. Wiltshire of The Martin Company in 1960 within the context of astrodynamics [31]. Although derived in the satellite-based Gaussian, or RSW coordinate system, the Clohessy–Wiltshire equations of motion (also called Hill’s Equations, or Hill–Clohessy–Wiltshire Equations) are typically mapped directly onto an xyz notation to reinforce and evoke the approximate nature of the solutions given. The Clohessy–Wiltshire equations of motion are

$$\ddot{x} - 2\omega\dot{y} - 3\omega^2x = f_x \quad (4.157)$$

$$\ddot{y} + 2\omega\dot{x} = f_y \quad (4.158)$$

$$\ddot{z} + \omega^2z = f_z \quad (4.159)$$

where the angular term $\omega \approx (d\theta/dt)$ is, by definition, the primary spacecraft’s mean motion. Note that whilst the relative vector, \mathbf{r}_{rel} , from the primary spacecraft to the secondary spacecraft was assumed to be small, the along-track range, y , does not appear in the Clohessy–Wiltshire equations of motion and is thus not directly restricted.

Alternative forms of the relative equations of motion can be found which are valid for a range of eccentricities; see [1, 32–34]. Further, significant errors, perhaps as much as 10 % for $e = 0.05$, can be found to accumulate for non-circular orbits when using the Clohessy–Wiltshire equations of motion.

To solve the Clohessy–Wiltshire equations of motion it is usual to assume that no external forces act on the secondary spacecraft, that is $\mathbf{F} = [0 \ 0 \ 0]$. This prohibits the analysis of continuous low-thrust maneuvers, but impulsive maneuvers can be treated by using the resultant velocities to define a new initial condition state. Noting the coupling between Eqs. 4.157 and 4.158 (in-plane motion), that Eq. 4.159 is a decoupled, simple harmonic oscillator (out-of-plane motion), the Clohessy–Wiltshire equations of motion can be solved using Laplace transformations as

$$x(t) = \left(4x_0 + \frac{2\dot{y}_0}{\omega}\right) + \frac{\dot{x}_0}{\omega} \sin(\omega t) - \left(3x_0 + \frac{2\dot{y}_0}{\omega}\right) \cos(\omega t) \quad (4.160)$$

$$y(t) = \left(y_0 - \frac{2\dot{x}_0}{\omega}\right) + \frac{2\dot{x}_0}{\omega} \cos(\omega t) + \left(6x_0 + \frac{4\dot{y}_0}{\omega}\right) \sin(\omega t) - (6\omega x_0 + 3\dot{y}_0)t \quad (4.161)$$

$$z(t) = z_0 \cos(\omega t) + \frac{\dot{z}_0}{\omega} \sin(\omega t) \quad (4.162)$$

where

$$\dot{x}(t) = (3\omega x_0 + 2\dot{y}_0) \sin(\omega t) + \dot{x}_0 \cos(\omega t) \quad (4.163)$$

$$\dot{y}(t) = (6\omega x_0 + 4\dot{y}_0) \cos(\omega t) - 2\dot{x}_0 \sin(\omega t) - (6\omega x_0 + 3\dot{y}_0) \quad (4.164)$$

$$\dot{z}(t) = \dot{z}_0 \cos(\omega t) - z_0\omega \sin(\omega t). \quad (4.165)$$

The application of the Clohessy–Wiltshire equations of motion will be discussed further in Chap. 12.

References

1. Griffin, M.D., French, J.R., “Space Vehicle Design, Second Edition”, AIAA, Reston, VA, 2004, pp. 103–192
2. Rivera, E.J., Laughlin, G., Butler, R.P., Vogt, S.S., Haghighipour, N., Meschiari, S., “The Lick-Carnegie Exoplanet Survey: A Uranus-mass Fourth Planet for GJ 876 in an Extrasolar Laplace Configuration”, *The Astrophysical Journal*, Vol. 719, No. 1, pp. 890–899, 2010. doi:10.1088/0004-637X/719/1/890
3. Roy, A.E., “Orbital Motion”, Taylor & Francis, Abingdon, Oxon, U.K., 2005.
4. Vallado, D. A., “Fundamentals of Astrodynamics and Applications”, Third Edition, Microcosm Press/Springer, Hawthorne, CA/New York, NY, 2007.
5. International Conference for the purpose of fixing “A Prime Meridian and A Universal Day”, *Protocols of the Proceedings*, Washington, D.C., U.S.A., Gibson Bros. Printers and Bookbinders, Washington, D.C., October 1884.
6. Cowell, P.H., Crommelin, A.C.D., “Jupiter, satellite VIII, the orbit of”, *Monthly Note of the Royal Astronomical Society*, Vol. 68, pp. 576, 1908.
7. Cowell, P.H., Crommelin, A.C.D., “Halley’s, perturbations of, in the past”, *Monthly Note of the Royal Astronomical Society*, Vol. 68, pp. 665–670, 1908.
8. Battin, R.H., “An Introduction to the Mathematics and Methods of Astrodynamics”, AIAA Educational Series, New York, 1987.
9. Bate, R.R., Mueller, D.D., White, J.E., “Fundamentals of Astrodynamics”, Dover Publications, Toronto, 1971.
10. Kyner, W.T., Bennet, M.M., “Modified Encke Special Perturbation”, *The Astronomical Journal*, Vol. 71, No. 7, pp. 579–582, 1966.
11. Stumpff, S., Weiss, E.H., “A Fast Method of Orbit Computation”, NASA Technical Note D-4470, April 1967.
12. Kaula, W.M., “Theory of Satellite Geodesy”, Blaisdell, Massachusetts, 1966.
13. Allan, R.R., *Proceedings of the Royal Society of London*, A228, Vol. 60, 1965.
14. Brouwer, D., “On The Accumulation of Errors in Numerical Integration”, *The Astronomical Journal*, Vol. 46, No. 16, pp. 149–153, 1937.
15. NASA Technical Memorandum, “US Standard Atmosphere”, NASA-TM-X-74335 [NOAA-S/T 76-1562], October 1976.
16. King-Hele, D., “Satellite Orbits in an Atmosphere”, Blackie and Sons Ltd., Glasgow, 1987.
17. Maxwell, J. C., “Electricity and Magnetism”, Oxford University Press, 1873.

18. Bartoli, A., "Il calorico raggiante e il secondo principio di termodinamica", *Nuovo Cimento* Vol. 15, pp. 196–202, 1876/1884.
19. Lewis, G.N., Letter to the editor of *Nature* magazine, Vol. 118, Part 2, pp. 874–875, December 1926.
20. Griffiths, D. J. "The Photon (1900–1924)." Section 1.2 "Introduction to Elementary Particles". New York: Wiley, pp. 14–17, 1987.
21. Carroll, B. W., Ostlie, D. A., "An Introduction to Modern Astrophysics", Pearson Addison-Wesley, 2007.
22. M^cInnes, C.R., "Solar Sailing: Technology, Dynamics and Mission Applications", Springer-Praxis, Chichester, 1999.
23. Macdonald, M., M^cInnes, C.R., "Solar Sail Science Mission Applications and Advancement", *Advances in Space Research*, Volume 48, Issue 11, pp. 1702–1716, December 2011.
24. Escobal, P.R., "Methods of Astrodynamics", John Wiley & Sons, Inc., 1968.
25. Tsien, H.S., "Take-off from Satellite Orbit", *Journal of the American Rocket Society*, Vol. 23, July-Aug., pp. 233–236, 1953.
26. Edelbaum, T.N., "Propulsion Requirements for Controllable Satellites", *ARS Journal*, Vol. 31, Aug. 1961, pp. 1079–1089.
27. Colasurdo, G., and Casalino, L., "Optimal Low-Thrust Maneuvers in Presence of Earth Shadow", *AIAA Paper* 2004-5087, Aug 2004.
28. Kechichian, J.A., "Low-Thrust Eccentricity-Constrained Orbit Raising", *Journal of Spacecraft and Rockets*, Vol. 35, No. 3, 1998, pp. 327–335.
29. Kluever, C.A., "Using Edelbaum's Method to Compute Low-Thrust Transfers with Earth Shadow Eclipses", *Journal of Guidance, Control, and Dynamics*, Vol. 34, No. 1, 2011, pp. 300–303.
30. Hill, G.W., "On the Part of the Motion of Lunar Perigee Which is a Function of the Mean Motions of the Sun and Moon". *Acta Mathematica*, Vol. 8, No. 1, pp 1–36, 1886.
31. Clohessy, W.H., Wiltshire, R.S., "Terminal Guidance System for Satellite Rendezvous", *Journal of Aerospace Sciences*, Vol. 27, pp. 653–658, 1960
32. Stern, R.G., "Interplanetary midcourse guidance analysis", Thesis (Sc. D.), Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, 1963.
33. Melton, R.G., "Time-Explicit Representation of Relative Motion Between Elliptical Orbits", *Journal of Guidance, Control and Dynamics*, Vol. 23, No. 4, July–August 2000, pp. 604–610.
34. Schaub, H., Junkins, J.L., "Analytical Mechanics of Aerospace Systems", Second Edition (AIAA Education Series), 2009.

Further Reading

35. Battin, R. H., "An Introduction to the Mathematics and Methods of Astrodynamics", American Institute of Aeronautics & Astronautics; Revised edition, Reston, Virginia, 1999.
36. Schaub, H., "Analytical Mechanics of Space Systems", 2nd Edition, American Institute of Aeronautics & Astronautics; Revised edition, Reston, Virginia, 2009.
37. Vallado, D. A., "Fundamentals of Astrodynamics and Applications", Third Edition, Microcosm Press/Springer, Hawthorne, CA/New York, NY, 2007.

Richard Brown, Tom Scanlon and Jason Reese

This chapter introduces the historical and technical background to the technology that is required for entry, whether from orbit or beyond, into the atmosphere of the various celestial bodies within our solar system. It opens with a historical description of the forms of atmospheric entry technology that have been employed on various missions to date. The physical constraints on the design of atmospheric entry vehicles are then described in broad terms. The atmospheric properties of the various near-Earth planets are then described: these properties have distinct bearing on the type of technology that is appropriate for any particular mission. Following a discussion of the range of modern computational techniques that exist for dealing with the very difficult problem of simulating the behavior of atmospheric entry vehicles, the chapter closes with a short perspective on future developments in the field.

5.1 Short History of Missions

5.1.1 Mercury

Although several probes have been sent on flyby missions, and the NASA MESSENGER probe is currently in orbit, no craft to date has landed on Mercury's surface. It is of note that the envisaged lander element of the ESA/JAXA Bepi-Colombo mission was removed early on in the mission design cycle due to the evident significant difficulties of that task. Indeed it is indicative of the difficulties of landing on a

planet with negligible atmosphere that, as envisioned, chemical thrusters would have been required during the descent sequence to provide the requisite deceleration for this probe to have reached the surface intact.

5.1.2 Venus

The Soviet Venera-3 (Russian: Венера-3) holds the honor of being the first probe to impact on the surface of another planet. This occurred on March 1, 1966. Unfortunately, as a result of a catastrophic early failure of its data capture system, the probe failed to provide any information during its descent. The subsequent Venera-4, -5 and -6 missions (1967–1969) survived down to a short distance from the surface of Venus, falling silent at ~ 20 km altitude, and managed to transmit valuable quantities of measured data before being crushed by the pressure of the Venusian atmosphere. Given the experiences with the earlier probes in the series, Venera-7 was substantially redesigned to withstand the pressures of 75–100 Earth atmospheres, which it was realized it would be subjected to on the surface of Venus. Despite a last-second failure of its parachute system, on December 15, 1970 it became the first vehicle to land on the surface of another planet. The probe continued to transmit information from the Venusian surface for 23 min before exhausting its batteries. Venera-8 to -14 (1972–1982) were all more or less successful, and returned significant amounts of atmospheric data as well as seismic and spectrographic information regarding the planet and its composition. In 1985, the Vega-1 and -2 probes culminated the Soviet exploration of Venus by deploying balloon-suspended instrumentation packages to explore the Venusian atmosphere. These devices floated in the atmosphere for about 46 h and revealed the clouds in the most active layer about 54 km above the surface to contain significant quantities of acid-rain, producing sulphuric acid. Although the exploration of Venus has largely been a Soviet affair, the US Pioneer

R. Brown (✉) · T. Scanlon
Department of Mechanical and Aerospace Engineering,
University of Strathclyde, Glasgow, Scotland
e-mail: richard.brown@strath.ac.uk

J. Reese
School of Engineering, University of Edinburgh, Edinburgh,
Scotland

Venus Multi-probe mission (1978) has also made a significant contribution to understanding of the Venusian atmosphere. As part of the mission, four atmospheric probes were targeted individually at the equator, the higher latitudes, and at the day and night sides of the planet—the day-side probe reached the surface and continued to transmit data for over an hour. Three of the probes were equipped with a nephelometer and temperature, pressure, and acceleration sensors, as well as a radiometer to map the distribution of radiative energy in the atmosphere. The larger, equatorial probe was equipped, in addition, with spectrometers to measure the atmospheric composition. The Doppler shift of the radio signals from all four probes were also used to characterize the winds and turbulence levels in the atmosphere.

5.1.3 Mars

Missions to the surface of Mars have been bedeviled by a long history of failures—if anything this is an indication of how difficult the Martian environment is to penetrate. The difficulties are well expressed by Braun and Manning [1] who show that, in strong contrast to the situation on Venus, the principal difficulty is to define a trajectory through the thin Martian atmosphere that will provide sufficient aerodynamic deceleration of the entry vehicle before the inevitable contact with the planetary surface. To date all successful missions have involved entry vehicles that have been sufficiently lightweight to meet very stringent bounds on their allowable ballistic coefficient, and all have been forced to be targeted at landing sites that were well below the mean surface level of the planet in order to exploit ‘more’ atmosphere. The first mission that was intended to enter the Martian atmosphere was the Soviet Mars 1962B probe—this mission failed before leaving Earth orbit. A dozen or so subsequent missions experienced various failures that prevented them from reaching Mars. In the early 1970s, however, the Soviet Mars-3 Lander (1971) and Mars-6 Lander (1974) managed to enter the Martian atmosphere but lost contact shortly thereafter. The NASA Viking-1 mission in 1976 achieved the first successful landing, returning images from the surface of planet, followed by the Viking-2 mission later in the same year. The Mars Pathfinder mission in 1997, followed by the Mars Exploration Rover mission in 2003 and the Mars Science Laboratory mission in 2012, all successfully landed rovers on the Red Planet. NASA’s Phoenix mission resulted in the first successful landing of a probe in the polar regions of Mars in 2008, with the aim of finding evidence of water and exploring the possibility of microbial life. Finally, there are ambitious long-term plans to send humans to Mars, together with the infrastructure that is required to support exploration of the surface for months or even years. The task of

landing the habitation modules (currently envisaged as having a mass of around 40 tons each) on the surface of Mars will pose significant challenges to the current generation of atmospheric penetration technology, and is stimulating research into novel heat shield materials as well as into the design of very large decelerators that will be stowable during transit and deploy only on entry into the Martian atmosphere.

5.1.4 The Gas Planets

In July 1995 an atmospheric entry vehicle was detached from the NASA/DLR Galileo spacecraft as it neared the Jovian system. The probe experienced accelerations of up to 230 G, and its warhead-style ablative heat shield was estimated to have lost about 80 kg of its 152 kg total mass during its subsequent passage into the atmosphere of Jupiter. Suspended by a parachute, the probe continued to transmit data for 58 min during which it descended 150 km into the atmosphere before failing once the ambient pressure reached 23 Earth atmospheres. Finally, on September 21, 2003, after 14 years in space and 8 years of observation of Jupiter and its moons, the mission of the Galileo spacecraft itself was brought to a close by sending it into Jupiter’s atmosphere. The primary reason for this terminal maneuver, though, was to avoid any chance of the craft eventually colliding with one of the Jovian moons and contaminating it with bacteria from Earth, and no useful scientific data was gathered during the process.

No probe has ever entered the atmosphere of Saturn. On December 25, 2004, however, the ESA Huygens probe, part of the joint NASA/ESA Cassini-Huygens mission, entered the atmosphere of Saturn’s largest moon Titan—the only natural satellite in the solar system known to have a dense atmosphere. While parachuting down to the surface, the probe relayed data regarding the physical and chemical composition of the atmosphere, as well as photographs of the surface of the satellite. The atmospheres of the two furthest planets from the Sun, Neptune and Uranus, have also yet to be penetrated by spacecraft.

5.1.5 Earth Atmosphere Reentry

During the second half of the 20th century, ballistic missiles and the associated reentry vehicle technology that was required to protect their thermonuclear warhead(s) during the final stages of an attack were vigorously developed on both sides of the so-called Iron Curtain. The technology associated with these entry vehicles remains largely classified except for some of the very earliest designs—which these days can even be visited in museums. The earliest

civil reentry vehicle designs were derivatives of this technology. NASA's Project Mercury (1959–1963), Gemini (1962–1966) and early Apollo program (1966–1968), along with the Soviet Vostok (Russian: Восток) (1961–1963) and Voskhod (Russian: Восток) (1964, 1965) missions pioneered the return from near space during the early days of human space exploration, employing a combination of ablative and heat-sink technology. Later crewed missions to the Moon (Apollo 1968–1972) involved a far more energetic return but used a derivative of the ablative heat shield technology used in the earlier missions. The proposed NASA/ESA Mars Sample Return Mission would pose an interesting problem because of its even more energetic return than Apollo, but advances in ablator materials technology in the half-century or so since the first missions to the Moon have been dramatic. In that vein, also worthy of note is the JAXA Hayabusa spacecraft that rendezvoused with a small near-Earth asteroid, called 25143 Itokawa, in mid-September 2005. On 13 June 2010 a small reentry vehicle survived the return of the main craft into the Earth's atmosphere and was recovered, together with its cargo of a sample of the asteroid's surface material, from its landing site at Woomera, Australia. This followed NASA's very successful Stardust mission which returned debris collected during its passage through the coma of the comet Wild-2 to Earth on January 15, 2006. Its advanced Phenolic Impregnated Carbon Ablator (PICA) heat shield allowed the sample container to survive a peak deceleration of about 25 G and heating rates approximately 30 times those experienced by the Apollo reentry vehicle.

By far the most intense activity requiring reentry into the Earth's atmosphere from orbit was associated with the early generation of reconnaissance and surveillance satellites. Contrary to modern practice, where images are transmitted back to Earth, in the early days exposed film was returned to the surface contained within purpose-designed reentry capsules. It is of note that although this practice continued, in a limited form, beyond 2010, exact details within the public domain are limited. Some of the larger satellites would be launched with several of these capsules on board for the purpose of returning film at various stages during the mission. The US Corona system alone had 163 (declassified) recoveries each involving entry into the atmosphere and subsequent recovery of a film capsule. By far the largest proportion of human space activity involving the need for reentry into the Earth's atmosphere since the 1970s has been associated with crew transfer to and from the Salyut, Skylab, Mir, and International Space Station, ISS. Apart from the transfers to and from Mir and the ISS using the Space Shuttle (see later) all the reentry vehicles proposed for these missions are characterized by their blunt-body aerodynamics (see later). This philosophy relies on the generation of significant forces parallel to the direction of

motion as a result of aerodynamic drag to provide the braking during atmospheric entry and often, but not always, a significant proportion of the subsequent descent. In comparison, the transverse forces that are required to maneuver or extend the trajectory are kept relatively small (although not necessarily negligible or un-exploited).

The modifications to the characteristics of atmospheric reentry and descent that can be achieved by exploiting the capability of suitably shaped bodies to generate significant aerodynamic lift (also see later) have been explored in a range of studies (e.g. the US Air Force Dyna-Soar, ESA Hermes, and JAXA HOPE programs). A number of flight articles have been built and tested (e.g. the US PRIME—Precision Recovery Including Maneuvering Entry and ASSET—Aerothermodynamic Elastic Structural Systems Environmental Test, and the Soviet BOR series of sub-scale research vehicles), and the technology has been embodied operationally in the NASA Space Shuttle Orbiter (1981–2011), the Soviet Buran (which made a single orbital flight under autonomous control in November 1988), and the Boeing X-37 (also known as the X-37 Orbital Test Vehicle) operated by the United States Air Force for orbital space flight missions. Future applications of lifting-body reentry technology may be to the ESA/NASA Crew Return Vehicle (a development of 1970s lifting-body research) and JAXA's Hyflex, but progress has been sporadic at best and, given the vagaries of funding for the development of such technology, there have been many false starts. Notable mention should also be made of the Scaled Composites SpaceShipOne and subsequent series of suborbital vehicles, developed as part of a private venture to take passengers to the edge of space. Return to the earth of these vehicles is, thermally-speaking, relatively benign since the energy that needs to be dissipated is some two orders of magnitude smaller than that required to be dissipated by a craft that is returning from even low Earth orbit. Thus, for these suborbital vehicles, atmospheric reentry can be achieved using simple, relatively low-performance atmospheric penetration technology in which the regions of the vehicle that are potentially thermally-sensitive are painted with a thin layer of ablative material.

The difficulties in creating a vehicle that is capable of repeated lifting entry from orbit into the planetary atmosphere are ably demonstrated by the US Space Shuttle. The orbiter vehicle relied on a winged glide-return following reentry from Earth orbit at a high angle of attack. The intention that the vehicle be capable of multiple returns into Earth's atmosphere posed severe constraints on the design of the thermal protection system (TPS) for the orbiter vehicle. Reinforced carbon-carbon was used to cover the areas exposed to the greatest heating (the nose and wing leading edges), while innovative (but fragile) silica-based low-conductivity tiles were used in areas subjected to intermediate heat loads (principally on the underside of the

vehicle). Silica or Nomex blanket material covered heat-sensitive areas on the sides and upper rear fuselage. As operational experience with the vehicle was gained, the composition and distribution of the heat shield materials evolved, but the underlying structure was still of aluminum and is similar in design to that of an aircraft. The first return of the orbiter vehicle from orbit (STS-1 on April 12, 1981) was accompanied by significant concerns that a number of missing or damaged silica tiles on the lower-surface TPS might lead to loss of the vehicle on reentry. Fortunately catastrophe was averted and Columbia's two test pilots returned safely to Earth on that occasion. Twenty-three successful missions ensued until Challenger was lost on January 28, 1986, for reasons largely unrelated to the subject matter of this chapter. A further 87 successful missions followed return of the Shuttle to flight on September 29, 1988. Then, on January 16, 2003, Columbia's TPS was damaged by a fragment of foam insulation shed from the external fuel tank during launch. As it reentered the Earth's atmosphere 2 weeks later, high-temperature air penetrated through the TPS into the aluminum structure of the vehicle through a hole in the leading edge of the left wing, causing a catastrophic high-altitude breakup of the vehicle. Although another 22 successful flights eventually followed the loss of Columbia, the Shuttle system never achieved its projected operational capability and has since been retired. Indeed, the entire US launch capability has reverted to traditional expendable launch vehicle technology, and, apart from the classified X-37 series of craft, makes use of conventional low-lift reentry vehicle technology.

5.2 Justification of Interest in Reentry Aerodynamics

The initial conditions for atmospheric entry are provided either by the entry corridor that permits capture of the vehicle from an interplanetary trajectory or, more simply if the ascending craft is already in a stabilized orbit, simply by the dynamic state of the vehicle post the de-orbit burn. Entry into the planetary atmosphere must take place following a carefully controlled relationship between speed and altitude to maintain the structural loads on the various components of the vehicle (and of course the loads on the occupants if the vehicle is crewed) and its temperature within design tolerances. The various possible atmospheric entry modes are shown in Fig. 5.1.

The ballistic trajectory is employed by entry vehicles that are incapable of generating significant lift, and the point of eventual impact of the vehicle with the planetary surface is essentially determined by the vehicle's dynamic condition at entry into the atmosphere. The glide trajectory is available to vehicles that have been specifically designed to

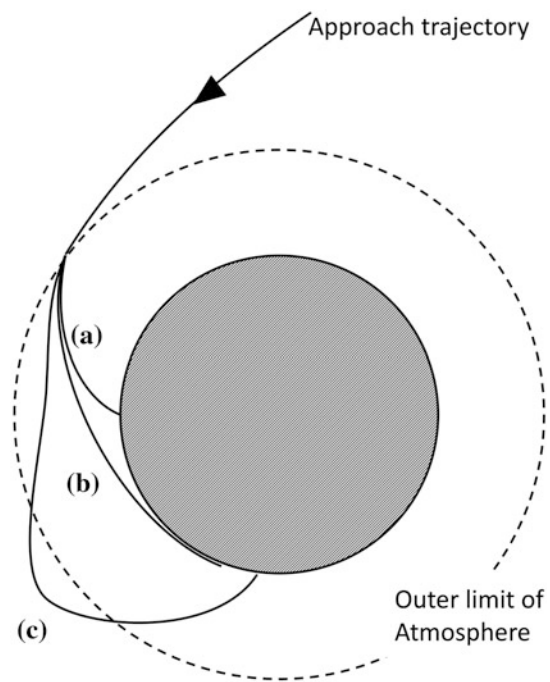


Fig. 5.1 The various possible atmospheric reentry modes. **a** Ballistic. **b** Glide **c** Skip

produce significant lift during entry, allowing the vehicle some flexibility in terms of the eventual landing site on the planetary surface. The skip trajectory allows a relatively gradual descent into the atmosphere via a series of relatively brief exposures to the rigors of the thermal environment associated with entry. Nevertheless, the integrated thermal load on the vehicle (see later) can be so great using this mode of entry as to preclude its use unless some method of rapidly dissipating the thermal energy from the vehicle can be arranged. Aerobraking maneuvers are an alternative means of achieving capture into a planetary orbit; in many ways this can be considered as a special case of the skip trajectory. See Chap. 4 for some further discussion of aerobraking. The entry trajectory of the vehicle is thus almost exclusively governed by its aerodynamic characteristics. The importance of the aerodynamic forces acting on an entry vehicle relative to gravitational forces is expressed crudely by the ballistic coefficient $B = m/(C_D S)$ for a non-lifting entry (where m is the vehicle mass, and S is the reference area for the vehicle's drag coefficient C_D), and by $L = B/(C_L/C_D)$ for an entry vehicle that is capable of generating significant lift (C_L is the vehicle's lift coefficient—see Fig. 5.2). The effect of these parameters on the subsequent trajectory of the entry vehicle is shown in Fig. 5.3. Vehicles with a low L or B tend to decelerate at higher altitudes within the atmosphere than those with low. This distinction is important when the aerodynamics of the vehicle is coupled to thermal characteristics of vehicle, as is described in more detail below.

Fig. 5.2 Force and moment coefficients as conventionally defined for an aerospace vehicle. All forces are non-dimensionalised by $1/2\rho(U_\infty)^2S$ and moments by $1/2\rho(U_\infty)^2Sl$

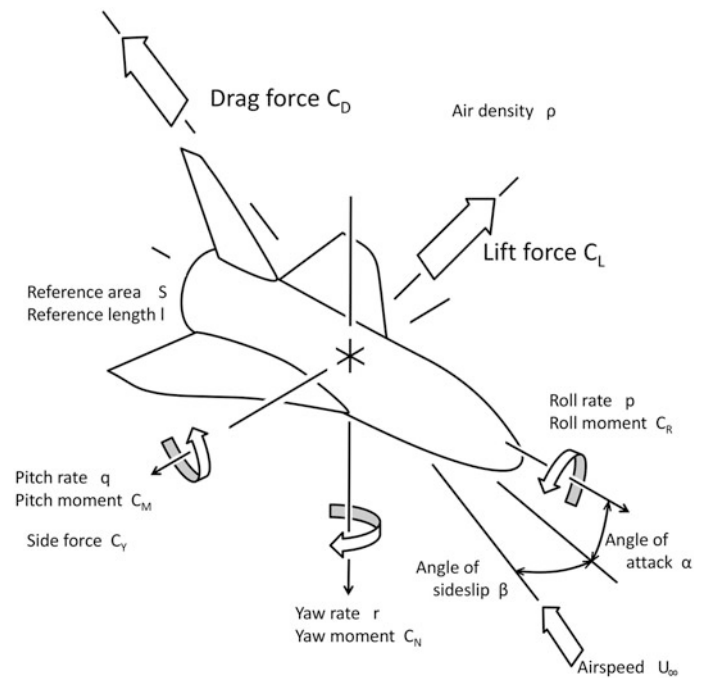
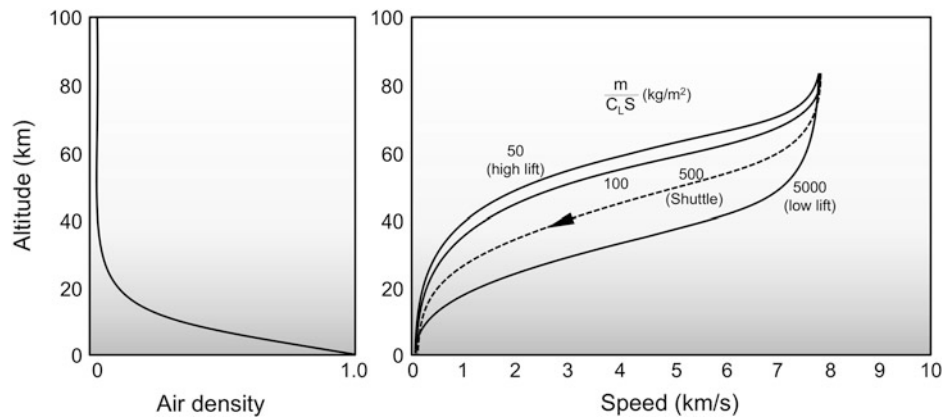


Fig. 5.3 Comparison between lifting and ballistic entry paths on the velocity-altitude map



5.2.1 Performance and Stability Implications

For lifting vehicles (even the blunt Apollo capsule was capable of generating useful force perpendicular to its direction of motion) the lift-to-drag ratio influences the variation in range that is achievable both along and transverse to the trajectory of the vehicle. In the hypersonic flight regime (specifically, when the Mach number is greater than about five), the lift-to-drag ratio of any aerodynamic body is relatively low. As shown in Fig. 5.4, values of C_L/C_D greater than 7 or 8 are difficult to achieve except possibly for waverider-type vehicles. Although the term ‘waverider’ has been misappropriated of late to apply to a broader range of hypersonic lifting vehicles, they remain largely theoretical constructs.

In practice Kuchemann’s correlation $(C_L/C_D)_{max} = 4(M + 3)/M$, shown in Fig. 5.4, is a useful rule of thumb but

it should be kept in mind that most practical vehicles have significantly lower C_L/C_D than predicted by this relationship. Even in the low-speed terminal flight phase, the lift-to-drag ratios of winged entry vehicles tend to be rather low given the requirement that the vehicles be slender and have relatively low aspect ratio wings in order to have good supersonic performance. For this reason, unpowered lifting vehicles such as the Space Shuttle Orbiter are capable of only relatively small lateral deviations, of the order of 2,000 km from their nominal entry trajectory (for instance as might be required to reach an alternate landing site). The integration of atmospheric-fed propulsion systems that would allow greater cross-range capability into the design of winged entry vehicles is an entire subject in its own right, however. For most entry vehicles, the drag coefficient is the most important parameter influencing their performance. Figure 5.5 shows the variation of this coefficient as a function of Mach number

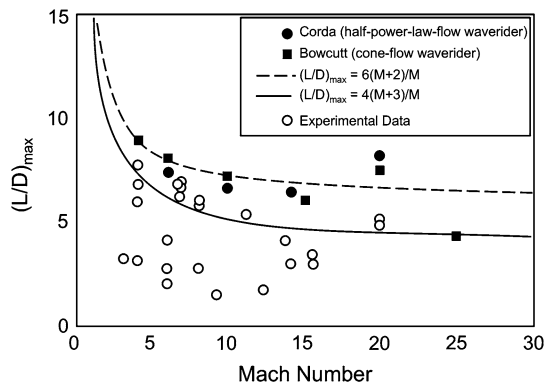


Fig. 5.4 Variation with Mach number of maximum lift to drag ratio of entry vehicles. *Open circles* represent historic test data [2]

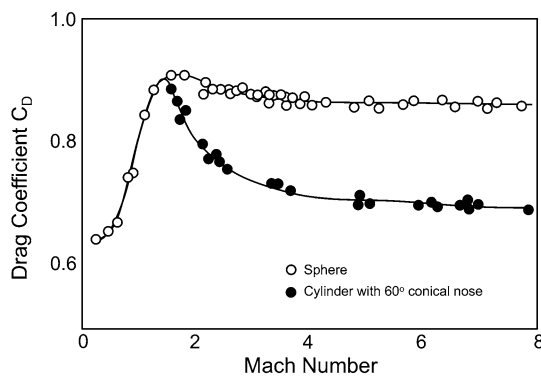


Fig. 5.5 Values of the drag coefficient for a sphere and for a cylinder with a conical nose [3]. The values do not change appreciably after $M = 4$

for a sphere and for a cylinder with a conical nose and suggests that, although the drag coefficient of any blunt, non-lifting entry vehicle would be dependent on its shape, it might not in fact be very sensitive to the actual Mach number of flight once in the hypersonic regime. This observation simplifies considerably the process of calculating the ballistic coefficient of the vehicle during those parts of the entry trajectory where the effects of heating are the most important.

The inherent stability of the vehicle during atmospheric entry is also a serious consideration. Without inherent stability an entry vehicle might tumble, thus interfering with later parachute deployment for instance, or adopt an attitude during entry in which its thermal protection system is ineffective. For all vehicles the first essential requirement for stability of the vehicle is that it should be able to attain an equilibrium attitude, in other words one in which all three components of the aerodynamic moment (pitch, roll and yaw) about the center of mass of the vehicle are zero. Stabilization of a non-lifting entry vehicle at a particular speed for a given altitude is usually assured by the fact that the increasing dynamic pressure associated with an

acceleration of the vehicle acts to increase the drag on the vehicle and hence to counteract the acceleration. The stability of non-lifting entry vehicles is thus primarily governed by the pitching and yawing moment derivatives with respect to attitude and rate of change of attitude; note that for an axially symmetric vehicle the pitching and yawing derivatives are generally very similar. Stability is generally assured if the (spring) derivatives $\partial C_M/\partial\alpha$ and $\partial C_N/\partial\beta$ and the (damping) derivatives $\partial C_M/\partial q$ and $\partial C_N/\partial r$ are selected to all lie in the design space that yields both a restorative moment and suppression of any oscillations following a perturbation of the vehicle from its equilibrium attitude (see Fig. 5.2). Stability in roll is often very difficult to achieve for entry vehicles that are close to being axially symmetric. This is because the roll moment derivatives are usually close to zero. This can have consequences for the design of parachute systems and retro-rockets intended for the final stages of descent, but can be exploited to allow spin-stabilization of the attitude of the vehicle. For a lifting vehicle the roll stability is affected by a similar pair of derivatives of the rolling moment, and there is a significant possibility of aerodynamic and inertial coupling between the roll, pitch and yaw degrees of freedom, especially given the slenderness of most lifting entry vehicles alluded to earlier. This coupling can yield an oscillatory instability that couples the yaw and roll of the vehicle into a mode known as Dutch Roll, or even to a rapid divergence in pitch or yaw attitude under certain flight conditions. Coupling of the dynamics of the vehicle with the atmospheric density fluctuations that are found at high altitude near the fringes of space can lead to an oscillatory instability in the coupled speed-pitch-altitude mode of the vehicle known as the Phugoid. There are significant contributions to all the aerodynamic forces and their derivatives from viscous, real gas and rarefaction effects, and these are not easy to estimate accurately during the design of a vehicle. Later in this chapter these predictive difficulties are discussed in more detail with application to the Stardust sample return capsule, and to the Shuttle Orbiter, where significant problems were encountered in matching practice with design as far as vehicle stability was concerned.

5.2.2 Thermal Implications

The aerodynamic and the thermal characteristics of the vehicle are closely coupled, particularly at the hypersonic speeds that are typical of the early stages of atmospheric entry. The intense friction generated at the surface of the vehicle is the prime source of heating, and the relative importance of the thermal energy imparted to the vehicle surface by friction, and the energy removed from, or imparted to, the surface by conduction in the gas is given by

Fig. 5.6 Velocity-altitude map, with superimposed zones in which various chemical or real-gas effects become important [4]

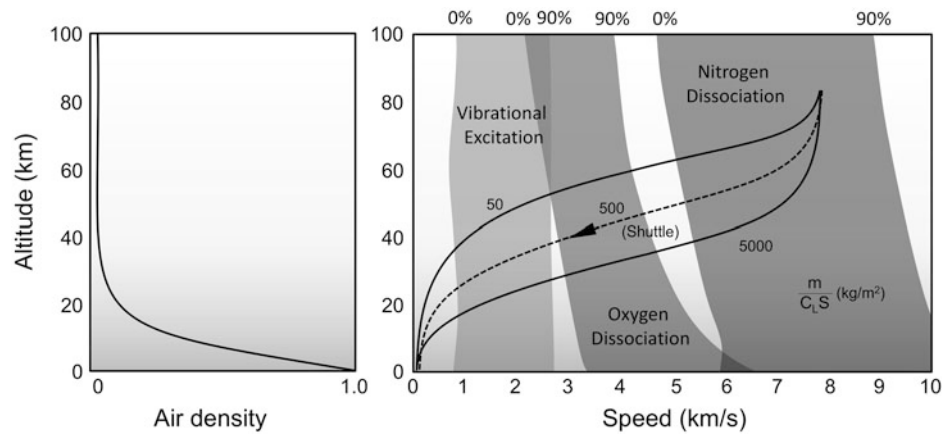
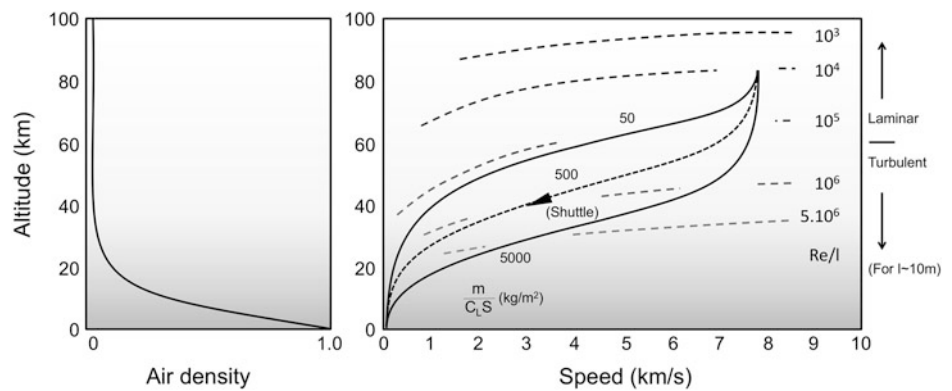


Fig. 5.7 Velocity-altitude map, with superimposed lines of constant unit Reynolds number (Re/l)



the Prandtl number $Pr = C_p \mu / k$ where C_p is the specific heat of the gas at constant pressure, μ is its dynamic viscosity and k is its thermal conductivity. As such the Prandtl number is principally a property of the gas, and for air has a value between 0.7 and 0.8. The deceleration of the gas in the boundary layer near the surface of the vehicle, and through the system of shock waves that is produced in the flow surrounding the vehicle results in very high gas temperatures. Under these conditions, it is quite possible that some of the energy of the gas molecules is then redistributed, through molecular collisions, from being purely kinetic (as in an ideal gas) to various forms of internal energy such as vibration and dissociation. Figure 5.6 shows the various chemical effects that can, in addition, be experienced by the vehicle at various heights and altitudes during entry to Earth's atmosphere. Changes in the chemical composition of the gas surrounding the entry vehicle have a significant effect on its thermodynamics, and thus on the thermal transfer to the vehicle, as well as on the distribution of forces on its surface.

The rate of heat transfer to the vehicle is also influenced strongly by whether the flow over the vehicle is laminar or turbulent. Laminar-turbulent transition is delayed at high altitude, as shown in Fig. 5.7, principally due to rarefied

flow effects. Indeed, the effective Reynolds number of the vehicle can be estimated fairly readily using the ideal gas relationship $Re = (M / Kn) \sqrt{\gamma \pi / 2}$ where M is the Mach number at which the vehicle is flying, γ is the ratio of specific heats of the gas in the flow around the vehicle, and the Knudsen number, Kn , introduced in Chap. 4 is the ratio between the molecular free path and the length-scale that best characterizes the vehicle (see the detailed discussion later in this chapter). This relationship needs to be treated with extreme care, however, since even at altitudes where the flow as predicted by this formula is ostensibly laminar, local protuberances (e.g. as caused by localized damage to the surface) can trip the boundary layer, creating zones of elevated heat transfer downstream; see Fig. 5.8 for instance, which shows a swath of excess heating on the underside of the Shuttle, resulting from local tripping of the boundary layer caused by localised roughness, possibly damage to one of the TPS tiles, on the forward starboard side of the vehicle. The likely location, even occurrence, of these micro-scale effects is notoriously difficult to predict and can force the designer to approach the specification of the vehicle's thermal protection system with a conservatism that is disproportionate compared to that which is applied to the remainder of the vehicle.

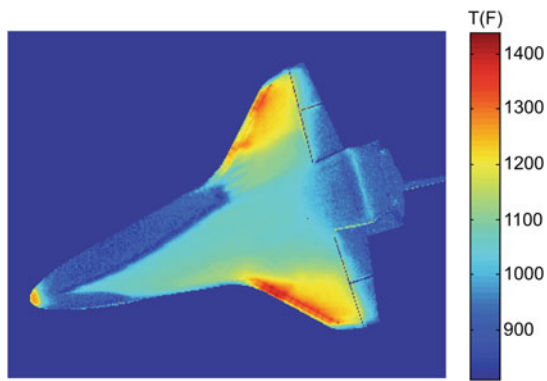


Fig. 5.8 Remote sensor image of the temperature distribution on the Shuttle lower surface during reentry into the Earth's atmosphere (*red* hottest, *blue* coldest). Note the asymmetry in the temperature distribution that results from boundary layer tripping, produced by localized roughness, possibly as a result of damage to one of the TPS tiles, on the forward starboard side of the vehicle (STS-134) [5]. *Image NASA*

Radiative heat transfer can also be a very important contributor to the thermal load on the vehicle, and is influenced by the surface condition of the vehicle, the state of the gas in which it is immersed (particularly with respect to its own radiative or absorptive properties), and whether or not other hot surfaces on the vehicle are exposed to the region under consideration via a direct optical path. For this reason concave portions of the vehicle (e.g. wing fuselage junctions, or the inwards-facing faces of fins) can be subject to much higher thermal loads than if just convective or conductive processes were involved.

The thermal transfer to the surface of the craft is affected very strongly by the surface roughness and chemistry, including the presence of catalytic reactions between the surface material and the constituents of the gas layer above the surface, and the presence near the surface of the chemical products of the decomposition of ablative shielding, should it be employed in the TPS of the craft. The Reynolds analogy is often exploited to relate the local heat transfer to the local skin friction. This empirical law holds that the rate of thermal transfer to the vehicle surface is directly proportional to the shear stress that the gas imposes on the surface, but such simple concepts break down at high altitudes and speeds as rarefied gas effects become more important (see later). The structure of the shock layer that precedes the vehicle at all supersonic Mach numbers has a profound effect on the heating rate at its forward extremities. This leads most entry vehicles to have a very blunt design in order to maximize the stand-off distance of the shock layer and thus to reduce the instantaneous thermal load on the vehicle. Even those entry vehicles designed to produce significant lift usually have blunt leading edges; this is simply a compromise between the performance of the vehicle (particularly in terms of lift-to-drag ratio) and the

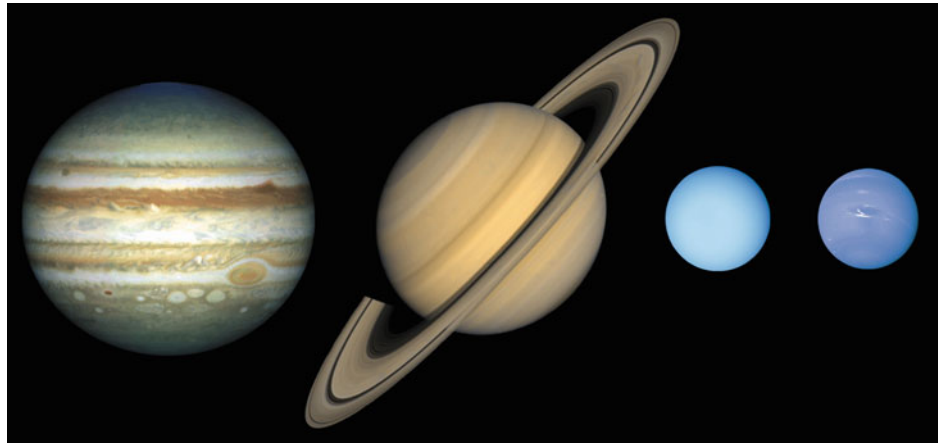
local heating of the structure. Again, as altitude and speed increase, real-gas (non-equilibrium) and rarefied gas effects can complicate considerably the prediction of the structure of the shock layer and thus the thermal transfer to the vehicle (see later). Trajectory design for entry into a planetary atmosphere is a careful balance between the maximum instantaneous thermal and mechanical load on the vehicle (which has implications for the mechanical resilience of the surface of the vehicle itself) against the total (integrated) thermal load that is transferred to the vehicle during its descent. The design process involves iteration between the design of the trajectory in terms of speed versus altitude (as governed by the aerodynamics and hence the shape of the vehicle), the type of thermal protection system that is used, and the design of the remainder of the structure of the vehicle. A variety of thermal protection strategies can be used, from the more usual passive or ablative systems to active systems where coolant material is ejected into the layer immediately adjacent to the surface or circulated underneath. The TPS must be carefully matched to the underlying structure of the vehicle; for instance, conduction from the thermal protection system into the underlying structure can cause the temperature within the vehicle to continue to rise even after landing, in the worst instances leading to eventual destruction of the contents of the craft. For this reason, and of course also to reduce the weight of the system at touchdown, many atmospheric probes have ejected their heat shield at some combination of speed and altitude before resorting to other means of slowing their descent to the planetary surface.

5.3 Characteristics of Planetary Upper Atmospheres

The characteristics of planetary atmospheres vary widely across the solar system. The atmosphere of each planet consists of a number of chemical species, the relative proportions of which vary with altitude. In this section only those planets closest to the Sun are considered in any depth. The closest three planets, namely Venus, Earth and Mars all consist of a rocky, solid, approximately spherical surface coated in a relatively thin layer of gas (atmosphere). On all these planets, radiative interaction with the Sun is strong enough to drive the atmospheric circulation that is primarily responsible for the distribution of weather (short-term variations in the atmospheric state, dominated by Coriolis forces) and climate (variations on the scale of the orbital period and longer) over latitude and altitude.

The ability of planets such as Venus, Earth and Mars to retain their atmospheres reflects a competition between the thermal velocity and the escape velocity of the gas molecules. Thermal energy causes some of the molecules at the

Fig. 5.9 The Jovian planets.
Image NASA



outer edge of a planet's atmosphere to have their velocity increased to the point where they can escape from the planet's gravitational field. The magnetic field of the Earth is strong enough to divert the solar wind (a stream of charged particles emanating from the Sun) away from the planet and thus to prevent it from stripping away the atmosphere. A schematic of the Earth's magnetic field can be found in [Chap. 3](#). Venus has no magnetic field, but interactions between the solar wind and the ionosphere of the planet protect its atmosphere to a significant extent. In sharp contrast, the magnetic field of the planet closest to the Sun, Mercury, is not strong enough to have prevented the atmosphere of that planet from being stripped away by the solar wind. As a result, Mercury does not have anything more than the most diffuse of atmospheres.

The planets furthest from the Sun (not including moons) are thought to be primarily gaseous in composition with their circulation being driven by internal processes that are not completely understood. Known as the Jovian planets, or gas giants, they consist of Jupiter, Saturn, Uranus and Neptune, as shown in [Fig. 5.9](#). In all cases, rapid planetary rotation plays a significant role in energy distribution and hence both weather and climate.

As discussed briefly in [Chap. 4](#), topography in certain geographical locations can rise to an appreciable portion of the atmospheric thickness resulting in large-scale modifications to the weather, due to deflection of winds and interruptions to circulatory patterns, e.g. the Himalayas on Earth and the Olympus Mons volcano on Mars as shown in [Fig. 5.10](#) at Earth.

5.3.1 Chemical Composition

The chemical composition of planetary atmospheres differs considerably. Venus has a very dense atmosphere with a surface pressure 90 times that at the Earth's surface and a chemical composition of 96.4 % CO₂ and 3.4 % N₂ with

trace amounts of SO₂ and water. The surface temperature is 740 K due to the greenhouse effect of the large quantity of CO₂. The atmospheric pressure on the surface of the Earth averages 101.325 kPa (1 atm), while the chemical composition by number is 78 % N₂ and 21 % O₂ with trace amounts of water vapor, argon, CO₂ and other gaseous molecules. The atmosphere on Mars is relatively thin (less than 1 % by mass of that of Earth) with a chemical composition of 95 % CO₂ and 2.7 % N₂ while the typical temperature ranges from 186 to 273 K. On Venus, Earth and Mars the pressure, density and temperature vary significantly with altitude. As discussed in [Chap. 4](#), numerous models exist for calculating these parameters, and best-practice guides exist to the various available atmospheric models for the Earth and the other planets, see for instance the AIAA Guide to Reference and Standard Atmosphere Models [6].

On Earth, the atmosphere is normally described in terms of a temperature variation as shown in [Fig. 5.11](#). Two equations for pressure are normally applied, one when the standard temperature lapse rate (the temperature gradient) is not equal to zero and one when it is equal to zero. [Figure 5.11](#) shows how the lapse rate varies with altitude while [Eqs. \(5.1\)](#) and [\(5.2\)](#) describe how pressure varies with altitude, with and without the lapse rate, respectively

$$P = P_b \left[\frac{T_b}{T_b + L_b(h - h_b)} \right]^{\frac{gM}{RL_b}} \quad (5.1)$$

$$P = P_b \exp \left[\frac{-gM(h - h_b)}{RT_b} \right] \quad (5.2)$$

where

P_b Static pressure (Pa)

T_b Temperature (K)

L_b Temperature lapse rate = -0.0065 K/m in International Standard Atmosphere [6]

h Height above sea level (m)

Fig. 5.10 Full-disk image of the Eastern hemisphere; the Himalayas can be seen to block cloud movement onto the Indian subcontinent. *Image* Indian Meteorological Department

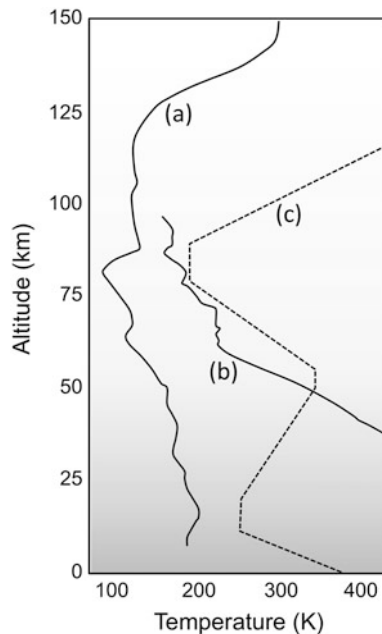
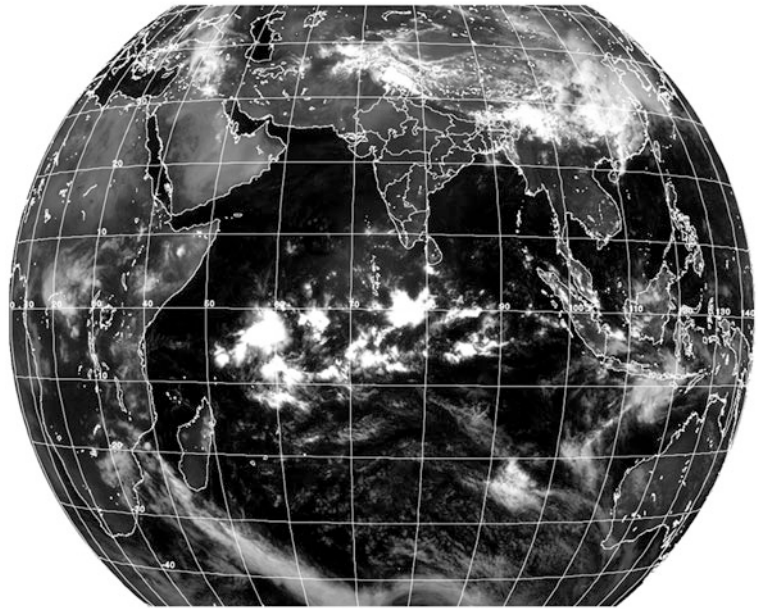


Fig. 5.11 Representative temperature variations with altitude for **a** Mars, **b** Venus and **c** Earth. Mars Pathfinder data (**a**), Magellan data (**b**), and International Standard Atmosphere (**c**)

- h_b Height at bottom of layer (e.g. stratosphere = 10 km)
 R Universal gas constant = 8.3142 kJ/kmol K
 g Gravitational acceleration (m/s^2)
 M Molar mass of Earth's air (28.9644 kg/kmol).

The mass density variations are almost identical, as shown in Eqs. 5.3 and 5.4 with and without the lapse rate, respectively

$$\rho = \rho_b \left[\frac{T_b + L_b(h - h_b)}{T_b} \right] \left(\frac{gM}{RL_b} \right)^{-1} \quad (5.3)$$

$$\rho = \rho_b \exp \left[\frac{-gM(h - h_b)}{RT_b} \right] \quad (5.4)$$

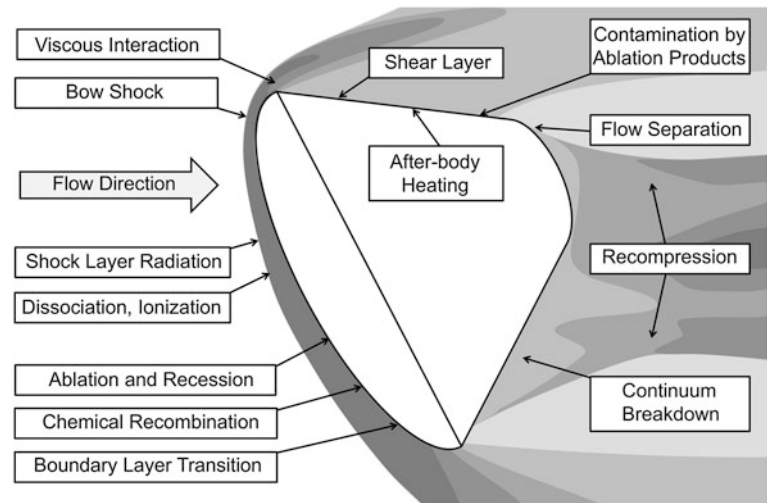
Figure 5.11 also shows how the temperature varies with altitude for Venus and Mars. This figure demonstrates how the temperature profile near the surface of all the planets is strongly influenced by convection, driven by radiative heating of the planetary surface by the Sun (exacerbated by the greenhouse effect in the case of Venus), while at higher altitudes direct absorption of radiation is the principal heating mode.

5.3.2 Weather

In general, the global circulation is driven by differential heating of the planetary surface at the poles and the equator. In the case of Earth and Mars, the orientation of their rotational axis to orbital plane creates seasonal weather patterns. Venus has an axial tilt of only 2.7° retrograde and this, combined with the exceedingly slow axial rotation and a CO_2 -rich environment, promotes an even distribution of temperature around the planet with little seasonal variation.

Winds at high altitudes are essentially geostrophic (a balance of Coriolis and pressure forces resulting in wind velocities that are more or less tangential to the isobars) whereas at lower altitudes the winds are affected by frictional contact with the planetary surface, generating an atmospheric boundary layer in which there is a significant

Fig. 5.12 Complex physics in the flow field surrounding a spacecraft during atmospheric entry. *Image NASA*



velocity component transverse to the isobars. Density variations are primarily driven by hydrostatic equilibrium but, at high altitudes, variations are large enough and of sufficiently long wavelength to interfere with spacecraft reentry trajectories, e.g. in the Shuttle Orbiter Approach and Landing Test, Free Flight 5, the wavelength of the vehicle's inherent Phugoid oscillatory motion was found to be of the same order as density perturbations in the upper atmosphere, leading to issues with aerodynamic stability [7].

5.4 Aerodynamics in the Upper Atmosphere

The flow surrounding a vehicle entering into a planetary atmosphere at high speed contains a range of complex physics, as shown in Fig. 5.12. Understanding the aerothermodynamics of atmospheric entry vehicles is a major challenge due to the combination of strong viscous effects, the possibility of shock/body and shock/shock interactions (all resulting in very high local heating rates), rarefaction phenomena associated with rapid variations in density through the flow, and the real-gas effects (vibrational excitation, dissociation and ionization) that result from the high temperature of the gas. Where an ablative TPS is used, additional complications are introduced into the gas chemistry and the properties of the flow near the surface of the vehicle, and indeed even the geometry and physical condition of the surface of the vehicle may be subject to rapid changes.

The design of atmospheric entry vehicles thus relies on three main approaches: (1) wind tunnel testing, (2) flight experiments, and (3) modeling and simulation. Historically, the wind tunnel has been the primary design tool, providing orders of magnitude more data than the other techniques combined. However, tests are costly and time consuming

and usually do not fully replicate real operating and flow conditions. Flight experiments deliver authentic data relating to some real operating conditions, although not necessarily those of most interest to the designers; capture of good data is difficult, and tests require meticulous and time consuming planning, at considerable expense.

The most promising route for future developments is through modeling and simulation, and the use of computational fluid dynamics (CFD), supported by wind tunnel testing where appropriate in order to validate and confirm the approach. The automotive and aircraft industries have already replaced the majority of their wind tunnel tests with CFD and the aerospace industry is fast following suit. While CFD has the potential to be relatively even more useful for aero-thermodynamic (high Mach number) flows than for low—because it can deliver data that cannot be measured or observed, under conditions that cannot be reproduced in a laboratory—it still faces major challenges, especially if the effects of the substantial variations in the air density or composition along the trajectory of the vehicle are to be fully characterized.

The use of the conventional Navier–Stokes–Fourier (NSF) model for aero-thermodynamic simulations requires that there is a strong separation between effects that occur on the microscopic scale and those on the macroscopic (bulk flow) scale. It is axiomatic that the NSF equations model the gas as a continuum in local thermodynamic near-equilibrium. In practice, this requires heat and momentum to be equilibrated almost instantaneously throughout the gas. For a monatomic gas, the molecules need some three or four collisions in order to equilibrate their (translational) energy and momentum with surrounding molecules. As, at normal temperatures and pressures, these molecules only travel on average a few tens of nanometers before they collide with another molecule, for most engineering systems operating at or near sea level the continuum description is

acceptable. Then linear stress/strain-rate (Newton's viscosity law) and heat-flux/temperature-gradient (Fourier's law) relationships can be assumed. However, at high altitude as the flow becomes more and more rarefied, the situation is not always so clear-cut.

5.4.1 The Knudsen Number and the Use of Bridging Functions for Rarefied Flows

The Knudsen number Kn is widely used as the decisive parameter to indicate the importance of rarefaction within a flow (as introduced in Chap. 4, classified into 'continuum', 'transition-continuum', or 'free-molecular' flow regimes). It is defined as the ratio of the mean free path of the gas molecules in the local freestream, to a characteristic system length-scale (such as the radius of curvature of the nose of the vehicle), i.e. $Kn = \lambda/L$ where $\lambda \propto \mu/\rho\sqrt{RT}$, with μ the gas dynamic viscosity, ρ its density, and T its temperature, all evaluated in the freestream. As the Knudsen number increases, the non-continuum, particulate-like behavior of the gas becomes ever more important. The continuum-fluid regime (in which conventional CFD modeling is applicable) holds up to $Kn \sim 0.1$, but for flows with $0.001 < Kn < 0.1$ (the 'slip regime'), some modifications of the surface boundary conditions in the fluid model are required in order to incorporate non-equilibrium effects. In the transition-continuum regime, $0.1 < Kn < 10$, the fluid acts neither as a continuum nor as free-molecular, although at higher Kn the flow can be described as free-molecular.

As an example, a reentry vehicle with a nose radius of around 10 cm encounters air at an altitude of 100 km with λ around 1 cm, which corresponds to $Kn \sim 0.1$. Shock waves from Mach 3 to 11 typically have $Kn \sim 0.2$ – 0.3 . These flows are therefore firmly in the transition-continuum regime. In most practical applications [8] a range of Knudsen number is encountered across different regions of the flow, e.g. high Kn in the vicinity of bow shocks, sensors or inlets, and moderate-to-high Kn in the wake or base flow downstream of the vehicle. Moreover, vehicles encounter a range of Knudsen number as they descend through the atmosphere owing to the variation in ambient air density. The aerodynamic modeler needs to ascertain the range of Knudsen number in the flow of interest, before choosing the most appropriate flow model (e.g. Navier–Stokes–Fourier, Direct-Simulation Monte Carlo, direct solution of the Boltzmann equation, etc.). The most modern aerodynamic methods are able to switch between models in response to these varying Kn conditions [9, 10].

Continuum-fluid flow predictions can be accessed through the NSF equations. Free-molecular flows require only the interaction of molecule fluxes with surfaces to be

accounted for, often yielding analytical expressions for the aerodynamic coefficients. The transition-continuum regime, however, poses great difficulties as neither the continuum fluid model nor the free-molecular model is valid. The difficulty in explicitly modeling this regime with accuracy has led to the widespread use of 'bridging functions'. These are correlations, many of which stem from experimental work in the 1960s and 1970s, that are used to interpolate values of aerodynamic coefficients between the free-molecular and continuum-fluid limits [11, 12]. Bridging functions are often used in aerodynamic analysis software to obtain transitional values of the axial force, normal force, and pitching moment coefficients, although other fluid parameters, such as the Stanton number (that measures the ratio of the heat transferred into the gas to its thermal capacity) can be modeled in this way too (see Fig. 5.13). However, it is important to realize that these functions are only approximations to the actual flow behavior, and estimates of their inaccuracy range between 5 and 20 % [10]. There is no substitute for full-field aerodynamic simulations of hypersonic vehicles across the Kn range, and this is increasingly possible due to the rapid development of high performance (parallel) computing.

5.4.2 A More Sensitive Indicator for High-speed Rarefied Flow Conditions

It is clear that the traditional definition of the Knudsen number results in a flow parameter that is not of the same kind as the Mach number and Reynolds number: in fact it indicates whether the molecular nature of the gas affects the thermodynamic characteristics of the flow. For the purposes of classifying atmospheric entry aerodynamics, with a freestream Mach number M_∞ and Reynolds number Re_∞ , a more sensitive indication of the flow rarefaction can be obtained through a term that quantifies, in non-dimensional form, the character of the molecular collision term in the Boltzmann equation [11], called the inverse Cheng parameter

$$K_c^{-1} = C^* M^2 / Re \quad (5.5)$$

where $C = \mu^* T / \mu T^*$, and $\mu^* = \mu(T^*)$ i.e. assuming a simple functional relationship between temperature and viscosity, and with T^* the simple average of the temperature immediately behind the normal bow shock and the temperature of the vehicle surface of interest. This parameter has been shown to correlate the flight and wind tunnel data for the axial force coefficient on the NASA Space Shuttle very effectively, for instance [12]. A proper analysis of which fluid models should be used in which ranges of K_c^{-1} is still awaited, however.

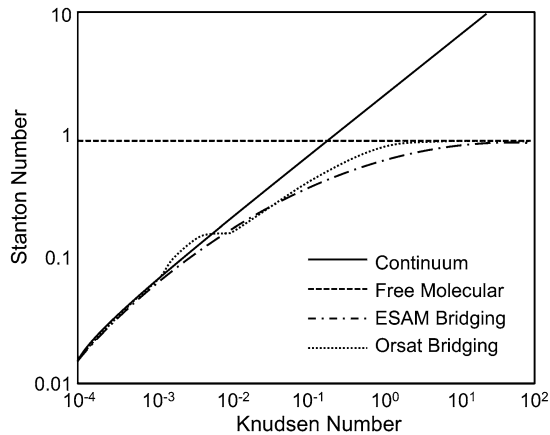


Fig. 5.13 Comparison of NASA’s Object Reentry Survival Analysis Tool (ORSAT) and ESA’s Spacecraft Entry Survival Analysis Module (SESAM) transition-continuum bridging functions for flow stagnation point Stanton number [9]

5.5 Consequences of the Incorrect Treatment of the Aero-thermodynamics

A rarefied flow is typically a thermodynamically, and often thermo-chemically, non-equilibrium flow. This is particularly the case in the context of high-speed atmospheric entry. There are several well-known examples where the non-equilibrium nature of the flow has strongly affected the aerodynamic performance of the vehicle, and two of these are presented here as an illustration of the attention to detail required in the aerodynamic modeling of entry vehicles.

5.5.1 Stardust

The Stardust atmospheric entry vehicle was part of a very ambitious NASA program that successfully returned samples, collected from the coma of the comet Wild-2, to Earth on January 15, 2006. An intensive program of computational analysis was conducted to support the eventual success of the mission, including free-molecular and Direct-Simulation Monte Carlo (DSMC) calculations of the capsule’s aerothermodynamics at the highest altitudes, the use of bridging functions in the transitional regime, and continuum Navier–Stokes in the lowest portions of the trajectory. These methods, and their application to the design and development of the Stardust capsule have been extensively documented in the technical literature, forming a valuable resource to aid future practitioners. One of the major concerns during the design was that the craft should remain in a stable attitude as it passed through the upper, rarefied parts of the atmosphere, so that its ablative heat shield would be presented properly to the oncoming flow during the time

over which the vehicle would experience its maximum heating. This was particularly important in this mission given that the PICA material used to construct the heat shield was being tested for the first time in flight, and that the heating regime for this mission was particularly severe. In fact the Stardust sample-return capsule was at the time the fastest man-made object ever to enter Earth’s atmosphere (traveling at 12.4 km/s or 28,000 mph at 135 km altitude).

The instability of planetary entry vehicles in the upper atmosphere (i.e. in the free-molecular and transition regimes) is not uncommon. In these regimes, surface shear stress accounts for more than 80 % of the forces experienced by the vehicle, and some components of this force can act to destabilize the vehicle [13]. Indeed, the Stardust capsule was calculated to be statically unstable for Knudsen numbers greater than 0.09, which, given the intended trajectory, implied aerodynamic instability for all altitudes above about 95 km. The capsule spin rate was thus designed so that the gyroscopic effect could be used to augment the stability of the capsule in the transition regime. It was realized however that too high a spin rate could lead to too much stiffness in pitch at lower altitudes, with the vehicle thus entering the regime of peak heating at too high an angle of attack. Selection of a suitable spin rate was complicated by the very small rotational inertia of the vehicle [14]. Even at lower altitudes, where the gas is better described as a continuum, the determination of the stability of the vehicle was complicated by real-gas effects, particularly by storage of energy in the vibrational modes of the gas, that significantly affected the position of the sonic line on the vehicle and hence the distribution of pressure on its surface. The changes in the shape of the heat shield that were a natural consequence of the use of ablative materials in its design also had an important effect on the evolution of the distribution of aerodynamic loads on the surface of the craft as it descended through the atmosphere, and had to be accounted for in detail in assessing the best overall configuration for the vehicle.

5.5.2 Space Shuttle Orbiter

In the high Mach number, high altitude part of its first reentry, the Space Shuttle Orbiter experienced a nose-up pitching moment that required the pilots to deflect the body-flap twice as much as they expected in order to restore trim (see Fig. 5.14). Explanations for this so-called ‘pitching moment anomaly’ included unaccounted-for compressibility, viscous, and real-gas (high temperature) effects. Compressibility and viscous effects were hypothesized because the low Reynolds numbers occurring in high-altitude flight mean that the body-flap might not have been effective when operating within a thick boundary layer. High viscous shear

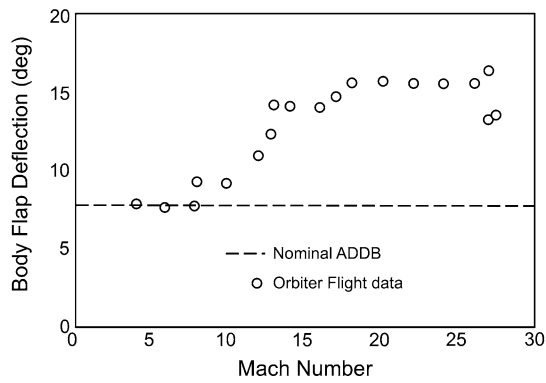


Fig. 5.14 The variation of the Shuttle Orbiter's actual body-flap deflection required for trim during its first flight [12], compared to the predicted value from the Shuttle's Aerodynamic Design Data Book (ADDB) [13]

of the cross flow in the nose region acting to induce nose-up pitching moments was also proposed. However, later numerical and experimental analysis [15] attributed the anomaly to high-temperature effects associated with the bow shock, leading to non-equilibrium and a lowering of the specific heat ratio of the gas flowing over the vehicle. This lower specific heat ratio caused pressures on the aft windward expansion surface of the Orbiter to be lower than was observed in the pre-flight hypersonic wind tunnel tests, given the inherent problems in matching true flight conditions, and in prior calculations which assumed ideal gas behavior. Small errors in predicting the distribution of pressure forces when integrated became a relatively significant error in the prediction of the pitching moment of the vehicle and thus on the deflection of the control surface that was required to restore equilibrium.

5.6 Simulating Aerodynamic Performance

5.6.1 Pre-supercomputer Methodologies

The accurate estimation of the performance of an entry vehicle and its thermal load, particularly in the hypersonic and supersonic flight regimes, has always been an important pre-requisite to successful mission design. In times prior to the advent of powerful computers, a variety of approximate methods were developed to achieve this task. Even today, these methods fulfill an important need that still exists for methods that can be used in the design context to estimate rapidly the aero-thermal characteristics of a given vehicle and to perform optimization and trade-off studies. The three main approaches that fall into this category are

- *Extrapolation from known data using hypersonic similarity laws*—The ideas of geometric and dynamic similarity that apply so well in subsonic and supersonic flows

can be readily extended into the hypersonic flow regime to allow the aerodynamic properties of a particular body to be extrapolated from the known properties of another. This process is effective as long as the two bodies are related by the correct similarity parameters for the problem at hand.

- *Component build-up methods*—In this class of approach, the geometry of the vehicle is decomposed into an amalgamation of simpler shapes, for instance spheres, cones and wedges, for which the flow properties can be calculated using simple theory or empirical data. Historically the most well-developed and readily available examples of this type of approach are the Douglas Aircraft Company series of Supersonic/Hypersonic Arbitrary Body Programs (SHABP), development of which started in the 1950s. In careful hands this approach can produce very reliable results, and indeed the preliminary design of the Space Shuttle was achieved using this type of technique.
- *Simplified gas models*—A range of simplified theoretical models for high-speed gas flow around aerospace vehicles have been based on the assumption that at the highest speeds, the random (thermal) motion of the gas particles can be neglected compared to their translation at the mean velocity of the flow. The particles of the gas are then modeled as traveling in straight lines before colliding with the surface of the vehicle. During a collision, a particle transfers momentum to the surface before rebounding, thus allowing the pressure on the surface to be estimated. These so-called Newtonian methods can be enhanced in their accuracy by various means, including correlation with experimental data and augmentation with empirical correlations, to yield surprisingly effective predictions, despite their inherent simplicity, of the aeromechanical and aero-thermal characteristics of modern entry vehicles.

5.6.2 Modern Computational Methodologies

Although it is known that the gas in the atmosphere is composed, at microscopic level, of discrete particles, a useful approximation arises if the particulate nature of the gas can be suppressed and instead the atmosphere can be treated as a continuum. Indeed, the continuum approach, particularly that through exploiting the Navier–Stokes–Fourier equations, is at the root of many very successful approaches to modeling the gas flow around aircraft and spacecraft, and yields good correlations with measured data over a wide range of practically-relevant operational conditions [16].

The problem that the continuum Navier–Stokes–Fourier equations have in capturing the properties of a rarefied flow, however, are highlighted in Fig. 5.15. The normal shock is a fundamental component of many high-speed aerodynamic

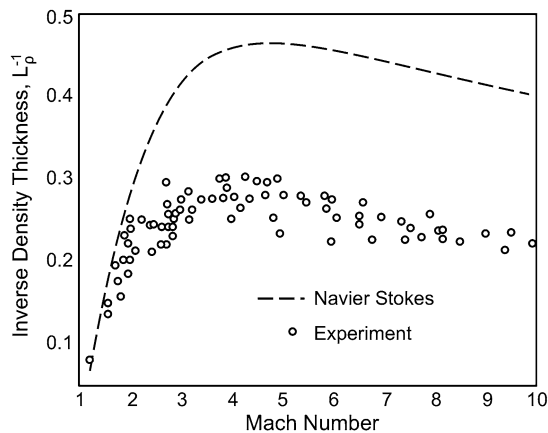


Fig. 5.15 The inverse of the density thickness of normal shock waves in argon gas up to Mach 11, measured experimentally and calculated using the conventional Navier–Stokes–Fourier equations [15]

flows, e.g. in the leading part of the bow shock in front of an atmospheric entry vehicle. Rarefaction causes the shock to be relatively thick (of the order of a few gas mean-free-paths, in other words of the same order as the stand-off distance of the bow shock from the vehicle leading edge). Figure 5.15 shows that the Navier–Stokes–Fourier (NSF) equations consistently predict shocks in argon gas to be about 50 % thinner than is observed experimentally. That the fluid dynamic equations have such difficulty in predicting the behavior in such a simple flow case, means that their validity in more complex rarefied flow-fields must also be called into question.

Non-continuum behavior can be accommodated to some extent in conventional continuum-based computational fluid dynamics (CFD) Navier–Stokes–Fourier approaches to modeling the gas dynamics of aerospace vehicles, for instance by incorporation of a finite slip velocity between the gas and any solid surfaces that are immersed within the flow [17]. A computationally efficient gas flow method, but one which has had only modest success to date, is to establish either a Kn -series or a Hermite polynomial approximation to the distribution function in the Boltzmann equation, which governs the gas behavior at particle-kinetic level. To first order (i.e. for near-equilibrium flows) both approaches yield the NSF set, but the solution methods can be continued to second and higher orders to incorporate more and more of the salient characteristics of a rarefied flow. This family of so-called extended hydrodynamic equations has various different members, including the Burnett, Grad 13-moment, R13, R26 equations, and others. Generally, however, these equation sets all have great difficulty in producing stable physical solutions for high-Mach-number flows. Their non-linearity also makes them difficult to solve numerically, and, as they are higher order in the gradients of flow properties, their solution requires additional boundary conditions that

are not easy to define. For these reasons, extended hydrodynamics has not established a firm place in the armory of tools that a high-speed aerodynamicist can deploy.

In highly rarefied environments ($Kn > 0.1$) accommodation of those non-continuum effects that occur in the flow away from surfaces remains a challenging problem. For this reason, analysis of gas flows in the non-continuum regime is most naturally conducted using specialized computational techniques that are derived from a statistical mechanical representation of the behavior of the individual particles comprising the flow. The most successful of these techniques is undoubtedly the Direct-Simulation Monte Carlo (DSMC) approach, originally proposed by Bird [18]. The DSMC technique allows particles to move and collide using kinetic theory considerations that treat the non-continuum, non-equilibrium gas behavior accurately. DSMC considers molecular collisions using stochastic rather than deterministic procedures, and each DSMC simulator particle represents a large number of real gas molecules. The decoupling of the particle ballistic motion from the physics that takes place during particle collisions improves the computational efficiency of DSMC greatly in comparison with other particle methods such as molecular dynamics (MD). As a result, the DSMC technique is becoming the dominant numerical approach for aerospace applications involving rarefied gas flows. Hypersonic vehicles which operate in rarefied gas environments may encounter chemically reacting flow conditions that can have a significant influence on aerodynamic performance and vehicle surface heat flux [19]. Numerical techniques which fail to incorporate such behavior miss out on an essential part of the flow physics surrounding the vehicle. One of the prime advantages of the DSMC method is the relative ease with which it can incorporate the real-gas, chemically non-equilibrium behavior of the flow. In the DSMC methodology, particle clusters can be endowed with the correct properties to capture both the kinetic and rotational modes of energy storage of the gas molecules. Taking the Earth as an example, vibrational excitation of the gas molecules as well as dissociation of both oxygen and nitrogen are likely to be important features of the flow around any hypersonic vehicle at the highest altitudes (80–120 km) and speeds, while even at lower speeds and altitudes vibrational excitation and limited dissociation of oxygen are still likely to be important [20]. Such real-gas effects need to be properly accounted for and normally the DSMC technique models dissociated and ionized air as a mixture of 11 species (N_2 , N_2^+ , O_2 , O_2^+ , NO , NO^+ , N , N^+ , O , O^+ , e^-), each described using appropriate DSMC molecular properties. In conventional CFD, reaction rates are calculated according to the Arrhenius law [20]. However, this assumes a local thermodynamic equilibrium distribution, which is inappropriate for rarefied hypersonic flows. In contrast, the

DSMC approach is able to capture the chemical reaction rates that actually apply under the non-equilibrium conditions that are typically encountered in high-speed rarefied gas flows. It should be realized though that even the DSMC formalism is an approximation to the full physics of the rarefied gas dynamics problem, and its inherent assumptions can limit its utility in practice. Eventually DSMC calculations may be supplanted by direct simulations, on a molecular level, of the flows around future atmospheric entry vehicles.

5.7 Future Perspectives

Although planetary exploration missions will continue to be founded, at least for the foreseeable future, on the basis of expendable, single-use technology, the Earth-return mission is likely to evolve in a different direction. Projected order-of-magnitude reductions in the cost of transferring payload to orbit renders reusable technology, founded on the principles of rapid turnaround and airline-like operations, an extremely attractive proposition. Initial steps have been made in the direction of embodying these concepts into flight articles, for instance the Virgin Galactic SpaceShipTwo suborbital vehicle (which undoubtedly will pave the way for more capable vehicles) and the Reaction Engines Skylon Single-Stage-to-Orbit vehicle, which is one of several configurations that are maturing on drawing boards around the world. The Space Shuttle taught that true reusability is difficult to achieve in practice, largely because of the very high maintenance required by current-generation technology, and future progress in this direction will require the development of low-maintenance, damage-tolerant, non-ablative thermal protection systems, robust propulsion systems that require minimal refurbishment between flights, and possibly new structural materials and fabrication concepts in order to achieve sensible payload mass fractions. Meanwhile, well-established and conservative design principles will continue to be used in the near term to provide a workable return capability from Earth orbit.

Proper design of future reentry vehicles will require the simultaneous consideration of many interlinking factors: calculations involving the coupled aerodynamics, structural dynamics and flight mechanics of any proposed vehicle will be necessary to avoid some of the mistakes of the past. The wide disparity in timescales between the various elements of fundamental underpinning physics will pose a significant challenge to computational techniques especially if the output is to be integrated seamlessly into the design process or into flight simulations.

A more accurate understanding of planetary atmospheres will be essential in order to de-risk future missions and to reduce the conservatism, and hence launch mass and propulsive requirements, of future planetary probes. It should

be acknowledged however that this is somewhat of a circular problem. Data from atmospheric probes provides only a sample of atmospheric behavior, usually over a very short timescale and at only one locality on the planet. Inference of the properties of the long-term behavior of a planetary climate, for instance on climatic timescales and synoptic or global length-scales, from this limited data is fraught with difficulties. Data fusion between accurate, localized information from atmospheric entry vehicles with data gleaned at long range from planetary orbiters is only part of the solution to ensuring sufficiently accurate predictions of the actual conditions that will be experienced by any probe as it penetrates the planetary atmosphere at the end of its long transit from Earth. The lack of any fully validated meteorological models for the atmospheres of our planetary neighbors remains a distinct element of risk to the success of any future mission.

The ideal of an all-encompassing numerical model that will allow the aerodynamics of an atmospheric entry vehicle to be predicted with high fidelity all the way from orbit, through the free molecular environment found at the upper fringes of the atmosphere, and down through the rarefied high-altitude layers to the essentially continuum-flow regime near the surface, still seems a long way from realization. At face value, the augmented Navier–Stokes–Fourier models would appear to offer the best hope, at least in capturing seamlessly the transition from rarefied to continuum conditions at intermediate and lower altitudes. Much effort has been devoted in the applied mathematics community to developing such techniques, but the ultimate goal is elusive. In any case, in the most highly rarefied parts of the atmosphere, a representation of the gas dynamics through a system of partial differential equations becomes entirely inappropriate, and a fundamentally different approach that acknowledges the particulate nature of the flow becomes essential. A robust thermo-chemical analysis tool is still lacking, and the traditional approach through pre-determining the set of reaction pathways that are available to the chemical constituents of the gas flow may eventually be supplanted entirely by a more fundamental approach that is founded on elementary quantum mechanics backed by relevant experimental data.

Integrating high fidelity aero-thermal calculations into the design process will be aided by future advances in computer hardware, including the use of graphical processor units (GPU) and cloud computing. The steadily reducing cost of high-performance computer hardware will most likely lead to a move away from concentrated national computational facilities to local or distributed hardware that is more accessible to a broader range of academic and industrial users. Indeed, the need for software that is more readily accessible to a wider range of users, and that is less dependent on the involvement of specialist practitioners,

will be key to robust spacecraft design in the future, particularly as it will allow broader oversight and verification of the design process. The open-source approach to software production offers a new and promising model for the integration of a broad range of technical contributions into a unified, verified approach to spacecraft aero-thermal design.

There will always remain a need for the output of any such software to be compared and checked against the behavior of the real world. As predictive methods become more accurate, the age and inadequacy of the present experimental database will become more and more apparent. Indeed, the most prolific data-gathering period was during the 1960s and 1970s when data assimilation and flow visualization techniques were in their infancy. Sadly, a major problem for the future validation of numerical techniques is posed by the increasing unaffordability, for many institutions, of high quality experimental equipment, and indeed the closure of many key facilities around the world during recent years. A strategic approach to the maintenance and expansion of experimental facilities, most likely on a national or even transnational scale, will be essential if the design of future atmospheric entry vehicles is to be properly supported by the next generation of numerical tools.

References

- Braun, R.D., and Manning, R.M., "Mars Entry, Descent and Landing Challenges," *Journal of Spacecraft and Rockets*, Vol. 44, No. 2, 2007, pp. 310-323.
- Anderson, J.D., Jr., *Introduction to Flight*, 4th ed., McGraw-Hill, Boston, 2000.
- Cox, R.N., and Crabtree, L.F., *Elements of Hypersonic Aerodynamics*, Academic Press, New York, (1965).
- Koppenwallner, G., "Fundamentals of Hypersonics: Aerodynamics and Heat Transfer." In: *Hypersonic Aerodynamics*, Von Karman Institute for Fluid Dynamics, Saint Genese, Belgium, 1984.
- Horvath, T.J., Zalameda, J.N., Wood, W.A., Berry, S.A., Schwartz, R.J., Dantowitz, R.F., Spisz, T.S., and Taylor J.C., "Global Infrared Observations of Roughness Induced Transition on the Space Shuttle Orbiter," RTO Applied Vehicle Technology Panel Specialists' Meeting, San Diego, USA, 2012.
- Anonymous, *Guide to Reference and Standard Atmosphere Models*. AIAA G-003C-2010, 2010.
- Powers, B.G., "Space Shuttle Longitudinal Landing Flying Qualities," *Journal of Guidance, Control and Dynamics*, Vol. 9, No. 5, 1986, pp. 566-572.
- Reese, J.M., Gallis, M.A., and Lockerby, D.A., "New directions in fluid dynamics: non-equilibrium aerodynamic and microsystem flows," *Philosophical Transactions of the Royal Society, Part A, Mathematical, Physical and Engineering Sciences*, Vol. 361, 2003, pp. 2967-2988.
- Lockerby D.A., Reese, J.M., and Struchtrup, H., "Switching criteria for hybrid rarefied gas flow solvers," *Proceedings of the Royal Society, Part A, Mathematical, Physical and Engineering Sciences*, Vol. 465, 2009, pp. 1581-1598.
- Wilmoth, R.G., Blanchard, R.C., and Moss, J.N., "Rarefied transitional bridging of blunt body aerodynamics," In: Brun R et al. (eds.) 21st International Symposium of Rarefied Gas Dynamics, Marseille, France, 1998.
- Macrossan, M.N., "Scaling parameters for hypersonic flow: correlation of sphere drag data," In: Ivanov MS, Rebrov AK (eds.), 25th International Symposium on Rarefied Gas Dynamics, St. Petersburg, Russia, 2007.
- Wilhite, A.W., Arrington, J.P., and McCandless, R.S., "Performance aerodynamics of aero-assisted orbital vehicles," AIAA Paper 84-0406, 1984.
- Mitcheltree, R.A., Wilmoth, R.G., Cheatwood, F.M., Brauckmann, G.J., and Greene, F.A., "Aerodynamics of Stardust Sample Return Capsule," *Journal of Spacecraft and Rockets*, Vol. 36, No. 3, 1999, pp. 429-435.
- Desai, P.N., Mitcheltree, R.A., and Cheatwood, F.McN, "Entry Trajectory Issues for the Stardust Sample Return Capsule," International Symposium on Atmospheric Reentry Vehicles and Systems, Arcachon, France, 1999.
- Muylaert, J., Walpot, L., Rostand, P., Rapuc, M., Brauckmann, G., Paulson, J., Trockmorton, D., and Weilmuenster, K., "Extrapolation from wind tunnel to flight: Shuttle Orbiter aerodynamics," AGARD-AR-319-Vol-2 Hypersonic Experimental and Computational Capability, Improvement and Validation, 1998.
- Bertin, J.J., and Cummings, R.M., "Critical Hypersonic Aerothermodynamic Phenomena," *Annual Review of Fluid Mechanics*, Vol. 38, 2006, pp. 129-157.
- Greenshields, C.J., and Reese, J.M., "The structure of shock waves as a test of Brenner's modifications to the Navier-Stokes equations," *Journal of Fluid Mechanics*, Vol. 580, 2007, pp. 407-429.
- Bird, G.A., *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford, 1994.
- Gallis, M.A., Bond, R.B., and Torczynski, J., "A kinetic-theory approach for computing chemical-reaction rates in upper-atmosphere hypersonic flows," *Journal of Chemical Physics*, Vol. 131, No. 12, 2009, pp. 124311/1-124311/13.
- Anderson, J.D., Jr., *Hypersonic and High Temperature gas Dynamics*, McGraw-Hill, New York, 1989.

Further Reading

- Hirschel, E.H., and Weiland, C., *Selected Aerothermodynamic Design Problems of Hypersonic Flight Vehicles*, Progress in Astronautics and Aeronautics Series, American Institute of Aeronautics and Astronautics, 2009.
- Murthy, T.K.S., *Computational Methods in Hypersonic Aerodynamics*, Fluid Mechanics and Its Applications, Springer, 2010.
- Park, C., *Nonequilibrium Hypersonic Aerothermodynamics*, John Wiley & Sons, 1990.
- Schweikart, L., and Hallion, R.P., *The Hypersonic Revolution: Case Studies in the History of Hypersonic Technology*, Vols. 1-3, Air Force History and Museums Program, U.S. Government Printing Office, 2003.
- Vinh, N.X., *Hypersonic and Planetary Entry Flight Mechanics*, University of Michigan Press, 1980.

David Alexander and Neil Murphy

The vast array of engineering, technology, manpower, and money required to prepare for and execute a spacecraft launch is focused on a single purpose, namely to put an operational payload in space. Often, the payload is simply regarded as the package to be delivered but the nature of this package, in particular its operational requirements, tend to drive the mission constraints as a whole—launch system, spacecraft, telecommunications, etc. Ideally, a well-designed mission would naturally start from the mission requirements whether they be scientific observations, Earth reconnaissance, or telecommunications, and build the system around those requirements. Typically, however, the payload design and sizing is constrained to fit a fixed budget, a given launch system, and all too frequently by the spacecraft to which it is to be attached.

The Merriam Webster Dictionary definition of a payload was introduced in [Chap. 2](#). NASA defines a payload in a very similar fashion, [1] as “*Any airborne or space equipment or material that is not an integral part of the carrier vehicle (i.e. is not part of the carrier aircraft, sounding rocket, expendable or recoverable launch vehicle). Included are items such as free-flying automated spacecraft, Space Shuttle payloads, Space Station payloads, Expendable Launch Vehicle payloads, flight hardware and instruments designed to conduct experiments and payload support equipment.*” This provides a working definition that is adopted in this chapter. Other chapters of this book address other subsystems defining the various components that lead to the presence of a specific piece of technology in space. However, it is difficult to fully appreciate the importance of

these subsystems on the design of a payload without making the connections where relevant.

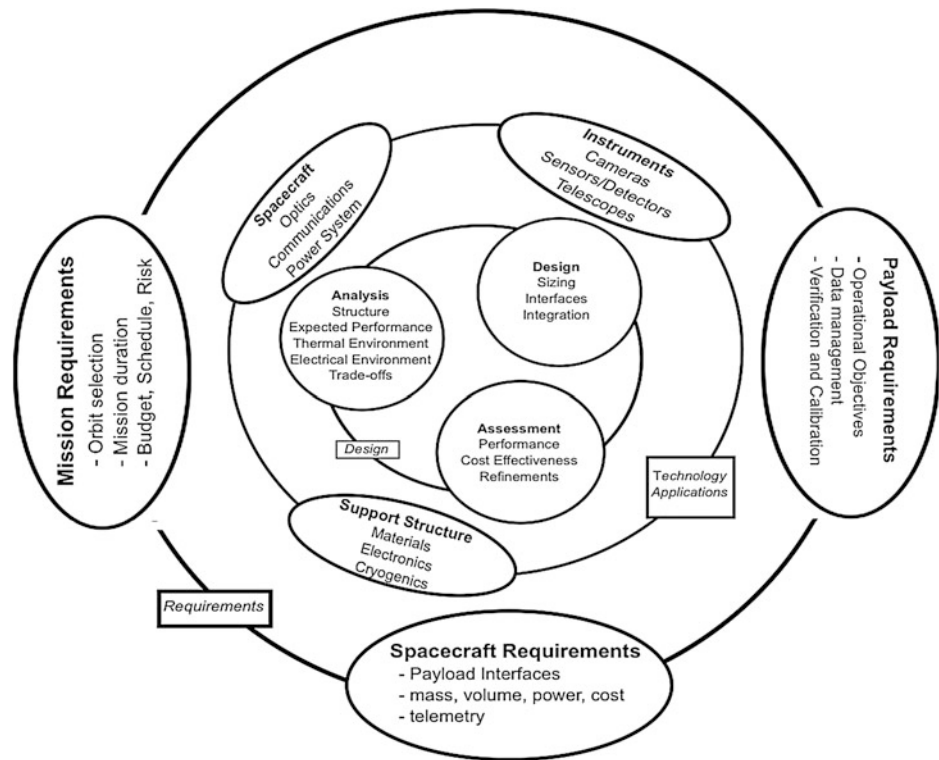
The payload must work with the spacecraft and launch systems to satisfy the requirements of the mission while complying with the various constraints imposed by these systems and by the overall cost and tolerance for risk. As such, there is a continuous interplay between the development of the payload, spacecraft, and overall mission. For example, the need for a payload to be in a particular orbit or trajectory has a major impact on the launch vehicle selection, the stability or pointing requirements of a payload instrument drive the necessary capabilities of the spacecraft, the day-to-day operation of a given payload has an impact on the optimal orbit selection and data storage and transmission capabilities of the payload, and so forth. The accommodation of the payload on a given spacecraft is defined by the various interfaces between the relevant subsystems. It is often the case that, at least for scientific payloads, a number of these interfaces are given in advance of the payload design, thereby guiding, and sometimes limiting, the capability of the payload. Optimizing the scientific or operational performance of the payload then requires a number of trade-offs. All of these issues will be discussed as this narrative develops.

There are a number of texts, most notably [2] and [3], that provide a step-by-step approach to designing a payload from the perspective of the detailed engineering of a space mission. This chapter takes a complementary approach by providing a comprehensive overview of the payload design process from an operational perspective, describing the range of factors that go into the definition of a payload, how it interacts with the various other subsystems that comprise a space mission, and how these factors impact the payload design. [Section 6.1](#) focuses on the basics of payload design, specific issues that are common to any and all payloads irrespective of their purpose. [Section 6.2](#) discusses mission resources and payload accommodation constraints that are

D. Alexander (✉)
Rice Space Institute, Rice University, Houston, TX, USA
e-mail: dalex@rice.edu

N. Murphy
Jet Propulsion Laboratory (JPL), California Institute of
Technology, Pasadena, CA, USA

Fig. 6.1 Design circle (after Fig. 2.1 in Ref. [2]) showing the interconnection between the different components and phases of the design process



externally imposed by, for example, the spacecraft and launch system. Section 6.3 deals with the design drivers, i.e. is directly related to what the payload is supposed to do, with Sect. 6.4 highlighting the constraints that impact but do not drive the design. Section 6.5 looks ahead to new technologies that will influence the kinds of payloads that will be flown in the future, and Sect. 6.6 provides some concluding remarks.

6.1 The Payload Design Process

Payload design is simply defined as the engineering approach that takes the various scientific or technical objectives of the mission, develops requirements for the measurements to be performed, and optimizes the performance of the payload within a set of specifications and constraints provided by the mission. More specifically, the design of a payload must be engineered to

1. Translate the operational requirements of the mission into a description of the payload performance and of the specific configuration needed to achieve this performance.
2. Develop a set of requirements on the spacecraft and mission as a whole that allow the payload to meet its requirements.
3. Assure compatibility of the various physical, functional, and program interfaces that couple the payload to the spacecraft and mission.

4. Achieve high confidence in the reliability, safety, and survivability of the payload in orbit.

Success in achieving these goals requires constant communication between the payload design and the various scientific, spacecraft and mission designs, and associated constraints and limitations. This requires an iterative process of design, analysis, testing and evaluation. Early in the design phase, these communications typically involve the exchange of high-fidelity mathematical models of the various structural, mechanical, electrical and thermal subsystems and their related interfaces. This is particularly important when the payload is large or complex, and can, consequently, have a significant impact on the spacecraft design. The application of such models early in the design is critical not only for optimizing the design but also for providing a benchmark against which the design can be verified. In addition, the mathematical models are extremely useful for getting an early start on the design and verification of the flight software.

One way to think of the payload design process is through a modified version of the conceptual design wheel of Hammond [2] as shown in Fig. 6.1. In this approach, the requirements bound the design process and drive the final payload configuration through the application of appropriate technologies with the concept being refined through an iterative combination of analysis, design definition, and assessment of the design, in relation to the bounding requirements. The analysis, design and assessment process may lead to a re-evaluation of the requirements, in some

Table 6.1 Payload categories with example payloads and missions

Payload category	Sub-category examples	Example payloads	Example missions and primary application
Scientific	Remote sensing	Cameras, imagers, spectrometers, telescopes	HST ^a , SDO ^b , IBEX ^c , WMAP ^d , AIM ^e
	<i>In situ</i> measurements	Magnetometers, radiometers, electric field measurements, ion spectrometers	Voyager [9], THEMIS ^f , Ulysses [11], Cluster [12]
	Sample return	Aerogels, gold foil collectors, dust collectors	Apollo 11 [13], Genesis [14], Stardust [15], Hayabusa [16]
	Planetary missions	Cameras, rovers, probes, radiometers, spectrometers	LRO ^g , Mars 1 [18], Cassini [19], Galileo [20], Viking [21]
Application	Communications	Transceivers, transmitters, transponders	Iridium [22], Intelsat [23], Spaceway-1 [24], Anik [25]
	Navigation	Transmitters and clocks	Inmarsat [26], GPS ^h , GLONASS ⁱ
	Reconnaissance	Cameras, infrared detectors, radar	SPOT ^j , KH series [30]
	Human space flight	Humans	Space Shuttle [31], ISS ^k , Soyuz [33]
	Space weather monitoring	Particles and fields measurements	ACE ^l , GOES ^m
Technology demonstration	In-space propulsion	Solar sails, plasma engines, ion drives, nuclear-electric	IKAROS [36], VASIMR ⁿ , DS1 ^o , JIMO ^p
	Disturbance reduction	Accelerometers, actuators	ST7 ^q
	Small sats	Test FPGAs, CMOS detectors	ST5 ^r , Cubesat [42]
	Formation flying	Metrology systems	Proba-3 ^s , PRISMA ^t

^a Hubble Space Telescope [4]; ^b Solar Dynamics Observatory [5]; ^c Interstellar Boundary Explorer [6]; ^d Wilkinson Microwave Anisotropy Probe [7]; ^e Aeronomy of Ice in the Mesosphere [8]; ^f Time History of Events and Macroscale Interactions during Substorms [10]; ^g Lunar Reconnaissance Orbiter [17]; ^h Global Positioning System [27]; ⁱ Global Navigation Satellite System [28]; ^j Système Probatoire d'Observation de la Terre [29]; ^k International Space Station [32]; ^l Advanced Composition Explorer [34]; ^m Geostationary Operational Environmental Satellites [35]; ⁿ Variable Specific Impulse Magnetoplasma Rocket [37]; ^o Deep Space 1 [38]; ^p Jupiter Icy Moons Orbiter [39]; ^q Space Technology 7 [40]; ^r Space Technology 5 [41]; ^s Project for On-Board Autonomy 3 [43]; ^t Prototype Research Instruments and Space Mission technology Advancement [44]

cases, or the need for different technologies. This is more likely when missions face significant constraints due to cost or launch energy considerations.

The technology component of Fig. 6.1 may seem initially to be an unnecessary complication as it could be argued that this might be a waste of resources and time compared to, say, designing the payload to the requirements with a fixed technology set as a limiting factor. However, as the design proceeds from one or more concepts to reality, the availability of the necessary technology, its maturity, or flight readiness, and the capability of the technology to meet the requirements and to meet them within cost and schedule are all critical to generating an optimal, robust, and capable design. Assessing the available technologies allows the payload design to take advantage of improvements in power consumption, better sensitivity, higher throughput and a whole host of other factors that might provide key discriminators between the original set of design concepts. The conventional wisdom is that about 80 % of the total-life cost of a mission is determined during the concept definition phase. In this light, technology assessment and selection becomes a major tool in driving the best design. Moreover, as

the design develops, the availability of alternative technologies provides a degree of flexibility in refining the design and in providing back-up possibilities in the event of increased performance, cost, or scheduling risk.

Payloads serve a number of purposes and can generally be segregated into three broad categories: scientific research, application, and technology demonstration. In some cases, payloads cover more than one of these categories; for example a payload could be launched for research purposes but be transitioned into an operational monitoring system, or perhaps a research payload component may also be demonstrating a new technology. Table 6.1 highlights these payload categories, breaking them down into sample sub-categories and giving examples of the sorts of payload components carried and missions that rely on these payloads. In addition to these broad categories of payload in terms of their desired outcome, such as scientific measurements or reconnaissance information, payloads within each category can be further classified in terms of the level of risk they entail relative to the importance of the payload to the commercial customer or agency responsible for its launch and operation. NASA has devised

Table 6.2 Classification considerations for NASA Class A–D payloads [1]

Characterization factor	Class A payload	Class B payload	Class C payload	Class D payload
Priority to agency	High priority, minimized risk	High priority, low risk	Medium priority, medium risk	Low priority, high risk
National significance	Very high	High	Medium	Low to medium
Complexity	Very high to high	High to medium	Medium to low	Medium to low
Mission lifetime	Long (>5 years)	Medium (2–5 years)	Short (<2 years)	Short (<2 years)
Cost	High	High to medium	Medium to low	Low
Launch constraints	Critical	Medium	Few	Few to none
In-flight maintenance	N/A	Not feasible or difficult	Maybe feasible	May be feasible and planned
Alternative or re-flight opportunities	No alternative or re-flight opportunities	Few or no alternative or re-flight opportunities	Some or few alternative or re-flight opportunities	Significant alternative or re-flight opportunities
Achievement of mission success criteria	Highest assurance standards used: minimum risk a focus	Stringent assurance only minor compromises: low risk maintained	Medium risk to mission success may be acceptable. Reduced assurance standards permitted	Medium or significant risk to mission success is allowed. Minimal assurance standards permitted

a four-class system, Classes A–D, which hierarchically defines each payload in terms of this risk-to-importance assessment. The factors that NASA have identified include criticality to the Agency Strategic Plan, national significance, availability of alternative research opportunities or reflight opportunities, success criteria, and magnitude of investment (Appendix A of Ref. [1]). The particular class in which the payload falls also factors into whether it is suitable for launch or operation on crewed platforms such as the ISS. Design constraints will be revisited later in Sects. 6.3 and 6.4 but by way of introduction to payload classification, the NASA payload classes are summarized in Table 6.2.

The payload is, of course, part of a larger mission and at a systems level is regarded as simply a component of the overall structure that is subject to all the necessary verification and testing required to assure mission success. However, while mission success requires that all subsystems and their interfaces perform nominally in space, it is the requirements of the payload design that most strongly drive the characteristics of the other subsystems.

6.1.1 Requirements Generation

The heart of good payload design, and indeed systems engineering in general as will be discussed in Chap. 7, is in understanding the top-level requirements and how they trace directly to the specific implementation of the payload, and how it interfaces with the spacecraft. The design process must always refer back to the requirements in order to

meet the design goals within the defined budget and schedule. Mission requirements vary widely depending on the nature and objectives of the mission and need to take into account a broad array of factors, including, for example, the required orbit, accommodation on the spacecraft, and operational objectives. These requirements have to be met while maintaining strict contamination control, surviving integration and testing, and considering the space environment in which the payload will operate. Much of the success of a payload will be determined by how well and how early these requirements are factored into the design.

In the previous section, three broad categories of payload were defined. As a subsystem, each category has many requirements in common with the others but significant differences appear when the specific nature of the payload is taken into consideration. For convenience, in this chapter, payload types are categorized as one of three types: scientific, commercial, or military. While the goals of these might be significantly different, with impact on the design process, the strict segregation is somewhat artificial since many space missions, regardless of their goals, have much in common. However, this simple categorization allows the impact of the mission goals on the design to be highlighted.

6.1.1.1 Scientific Research Payloads

The requirements unique to scientific payloads are primarily focused on meeting the observational goals of the mission within the given constraints of cost, power, mass, volume, reliability, operational lifetime, data collection rate, on-board data storage, and telemetry rate. Thus, the various components

of the observational system, e.g. telescope, camera, detectors, and data acquisition, have to be optimized in the design to perform to the required specifications, within the mission constraints. An additional design concern for some instruments is the control of electromagnetic interference (EMI) or magnetic contamination from other systems or other parts of the payload. Elements of the payload, such as magnetometers, or plasma wave instruments, will levy requirements on the system for acceptable levels of EMI and magnetic cleanliness and, at the system level, limits on contributions to the spacecraft EMI and magnetic environment are often levied on individual subsystems. An example of this might be limits on the external magnetic field from a subsystem. EMI and magnetic contamination are also mitigated by the use of booms to mount sensors. Other forms of contamination are also of concern to payload elements; for example, ultraviolet spectrometers and imagers can be sensitive to particulate, molecular, or chemical contaminations and infrared instruments can be sensitive to thermal contamination (requiring cooling of their optical elements). As payloads become more complex, for example with multiple cameras, greater attention must be paid to the design of the electronics, thermal control, and data control and management. Careful optimization of the payload configuration can reduce the number of electronics boxes, radiators and CPUs with concomitant impacts on the mass, power and volume budgets.

In addition, many scientific payloads have a range of required observational modes that are designed to optimize the science return. These modes can impact the design of the payload by increasing system complexity. For example, accommodating operational flexibility may require additional mechanisms (e.g. filter wheels), smart data control (e.g. on-board analysis and data selection), responsive changes to instrument operation (e.g. changing exposures, shutter controls, etc.), or ground-level command control (e.g. regular upload of observational command sequences). These are all important factors that drive the design.

6.1.1.2 Commercial Payloads

Commercial payload requirements are more concerned with meeting the goals of the customer that the payload is to serve and so schedule, robustness, safety, reliability, continuity, and communications, which are important for all payloads, take on an increased significance for commercial missions. While different commercial uses will require a different set of design criteria, the common thread between them all is to satisfy the above goals and to do so in a cost-efficient way and that frequently involves multiple launches of similar payloads. It was noted in the Commercial Space Transportation Study [45] performed by NASA in 1997 that, with respect to payloads, the design process must allow for enhanced system flexibility that facilitates a payload change-out up to 5 days prior to launch for the same

payload and 30 days pre-launch for a different payload. The payload integration has to be simplified to accommodate the focused (often single-purpose) operations, the standardization of the integration process (frequently commercial payloads are not unique as in the case of scientific payloads), and the more stringent schedule. These factors, when coupled to the business operations of the commercial space services provider, require repeatability in the payload design, a standard set of interfaces, a limited if not restricted set of operational modes, and compatibility with a standardized set of launch vehicle and spacecraft conditions and functions.

Until recently, the focus of the commercial space industry was the launching and operations of payloads for a range of commercial uses, such as navigation, telecommunications, and entertainment. However, recent trends have seen an increased emphasis on space tourism and, with the end of the Space Shuttle program, the servicing of the International Space Station. Thus, the primary objects being delivered to space would be humans or cargo, which primarily have a design impact on the spacecraft and launch vehicle and will be addressed later chapters.

6.1.1.3 Military Payloads

Crudely speaking the requirements for military payloads are unique in that they place a strong focus on the same criteria as commercial payloads while embodying the instrument uniqueness of a scientific payload. The most common functions of a military payload are reconnaissance, early warning capability, and space situational awareness, but the range of objectives also includes navigation, secure communications, and space-based weaponry. One unique feature of military space payloads is the relative importance of technology demonstration missions. By their nature such payloads tend to incur more risk than might be tolerated on an operational payload. Collection of engineering and performance data for subsequent analysis also impacts the design goals. Finally, an area of increased recent debate is the notion of a hosted payload where a government payload, including military and intelligence missions, is 'hosted' on a commercial launch vehicle. This necessarily impacts the payload design but also adds significant risk to the host payload and launch vehicle, most specifically matching requirements and impact on schedule.

6.1.2 Traceability of Operational Requirements

Generating the requirements for a given payload is the first and most important step in designing the payload to meet the mission goals. It is also important to have direct traceability of the payload design to the specific requirements that factor into that design. The characterization of this

Table 6.3 Generic scientific traceability matrix

Science objectives	Science measurement requirements	Instrument functional requirements	Top-level mission requirements
Science objective 1	Data format (images spectra)	Exposure times	Orbit selection
Science objective 2	Spatial resolution	Dynamic range	Mission duration
Science objective 3	Field-of-view	Wavelength selection	Operation mode (e.g. daily uploads)
	Temporal cadence	Wavelength range	Telemetry rate
	Exposure times	Time resolution	Data collection rate
	Time resolution	Spatial resolution	Data compression
	Required resolution of physical parameters of interest (temperature, magnetic field, density, etc.)	Spectral resolution	On-board memory
		Image stability	Ground station coverage
		Signal/Noise	Ground system management
		Effective area	Tracking and command control
			Ground data storage

process often takes the form of a requirements traceability matrix that documents the flowdown from the operational goals to the specific modes of operation of the payload and ultimately to mission requirements such as choice of orbit and required telemetry bandwidth. Each aspect of a given design must be able to trace its origin to a requirement or set of requirements and this leads to traceability matrices for each of the mission requirements, spacecraft requirements, payload requirements, and operational requirements. In all cases, the relevant traceability matrix identifies each goal of the mission, what it takes in terms of measurements to achieve that goal, how these measurements are attained by the particular device on the payload, and how the data collected is compressed, stored and downloaded. An example of a generic traceability matrix for a science payload is shown in Table 6.3.

Each science objective can be traced to a specific set of requirements to attain the science and a specific set of observations that will meet those requirements within the constraints of the mission requirements.

6.1.3 Accommodation Requirements

The payload, of course, is part of a larger system and, as a result of this, the mission and spacecraft impose constraints on the payload design. Ideally, the payload should be designed to satisfy the higher level requirements provided by the overall objectives of the mission. However, it is more common to optimize the design within significant constraints that cater to a specific cost cap, spacecraft bus, launch vehicle, or other resource budgets. The following section discusses how the resources available to the mission affect the payload. Here the focus centers on the spacecraft accommodation of the payload and how this is optimized.

Key issues in the accommodation of the payload with the spacecraft are the resources required by the payload, the

nature and number of the interfaces and the impact these have on the payload operation, and the nature and magnitude of the various ‘disturbances’ to the payload that are generated by the spacecraft and its environment. Much of this is discussed in other chapters but it is useful to highlight some of the key issues here.

6.1.4 The Spacecraft Environment

A number of the aspects of the spacecraft environment need to be considered as part of the payload design process. The launch environment produces significant mechanical, vibration and acoustic stresses. While this environment is transitory, it is one of the most significant drivers on the mechanical design of the payload. Demonstrating that a payload can withstand the launch environment, as transmitted via the spacecraft, is a key part of the flight qualification process. The mechanical design of the payload must take into consideration the spectrum and intensity of the vibration environment, including the potential for resonances. In addition, the low-frequency shock at liftoff, the static load during launch, and the high-frequency pyrotechnic shock at separation can threaten components at specific points on the spacecraft and in the payload. To avoid the risk of damage to sensitive components, a launch lock mechanism, released in orbit, is often required. Once in orbit, the mechanical environment is usually benign, although orbit insertion around another body can also produce significant mechanical stresses. While these are usually less severe than during launch, particular attention must be paid to payload elements that are deployed after launch, such as antennas or booms. The acceptable thermal environment of the payload can also be a significant driver on payload accommodation design. Payload thermal requirements are often given as allowable operating and non-operating (or survival) temperature limits, and not surprisingly, the wider the temperature range over

which the payload can operate and survive the simpler the thermal design and interfaces. Often there are particular parts of an instrument that need to be cooled, or have their temperatures precisely controlled, for example a CCD or a bolometer. In this case the payload is usually responsible for the cooling or temperature control, and the thermal requirements on the spacecraft could include the availability of space for a radiator, or a limit on the heat flow from the spacecraft to the payload. The use of active cooling (or heating) is another driver on payload design, and can be as simple as a thermoelectric cooler (TEC), which will require power and a heat sink, or a cryocooler, although this can impose vibrations on the system during operation.

6.1.5 Payload Interfaces

All possible interfaces associated with a payload and payload-spacecraft pairing are defined in the Interface Control Document (ICD), which will itself be discussed further in [Chap. 7](#). A number of interfaces can be identified, including payload-to-spacecraft interfaces, inter-payload interfaces (say between an instrument's optical path and the structural design) and space-to-ground system interfaces. In addition, other mission-level requirements, such as controlling and/or mitigating the level of contamination, may have consequences for the interface control (see [Sect. 6.1.9](#)). Many of these are discussed in later chapters. In this chapter the focus is on the payload-spacecraft interfaces.

The payload interfaces consist of the mechanical, thermal, and electrical connections between elements of the payload and the spacecraft. The mechanical interface attaches instruments to the spacecraft, maintains alignment of the instrument to some specified tolerance, and transmits launch loads and stresses to the instruments. It also forms part of the thermal interface. The thermal interface couples the transfer of thermal energy to and from the payload, either between the payload and spacecraft, or the payload and the external environment, for example via radiators. As part of this interface, the spacecraft may provide heaters and temperature monitors to maintain the payload temperatures within the allowable survival range in the event of payload power being removed. The electrical interface carries power and signals between the payload and the spacecraft system. Requirements that drive this interface include the required power for the payload, the data transfer rates, and the necessary housekeeping data. Payload communication is often carried out over standard interfaces, such as RS422 or Spacewire, which have a wide range of capabilities and substantial flight heritage. Power transfer is also often standardized. Many spacecraft buses provide power at a nominal voltage of ± 28 V—in such cases, the payload will provide conditioned power at voltages required by the

payload electronics, for example field programmable gate arrays (FPGA) often need a 3.3 V supply. In some instances, the spacecraft provides conditioned power at the necessary voltages, with the required noise performance. This can save mass by consolidating power supplies, but may also increase system complexity and may still require power conditioning in the payload. It should be noted that a failure of the power system can essentially signify the death of a payload. Although consolidating power sources may provide some efficiencies, these should not come at the expense of reliability. Interface definitions need to be developed early in the payload design process, as they are an important tool in understanding the impact of the payload requirements on the system as a whole and the constraints placed on the payload by the spacecraft.

The repeatability, sustainability, and affordability aspects of commercial and military missions, with multiple spacecraft being commissioned with similar operational goals, pushes the definition of standard interfaces and integration procedures. It is possible, in some cases, to modify the interfaces but this typically comes with increased cost, complexity, time and risk. Most missions are constrained to fly on a given launch vehicle with a given spacecraft architecture and as such many of the interface specifications are preset and allow little flexibility in the design. In some cases, this can limit the ability to optimize the payload to meet the mission requirements, although modern interface protocols have a wide range of capabilities.

6.1.6 Payload Integration

On a conventional spacecraft, the integration of the payload, spacecraft, and launch vehicle can take anywhere from 18 months to 3 years or more, depending upon the complexity of the payload (see [Fig. 6.2](#)). Some integration processes can be quite complex with the payload requiring a series of additional operations, such as purging, vacuum pumping, or instrument cooling, during the integration process. During the integration and subsequent testing, the payload is attached to the spacecraft, the functionality of the interfaces is tested and compared to predictions, and then the performance of the system is tested and compliance with requirements is established. Final tests are carried out after integration to ensure that the entire system is capable of achieving its required performance in the intended space environment, and that it will survive the rigors of the launch.

6.1.7 Orbit Requirements

For the majority of space missions the chosen orbit is a fundamental component of the mission, enabling, for

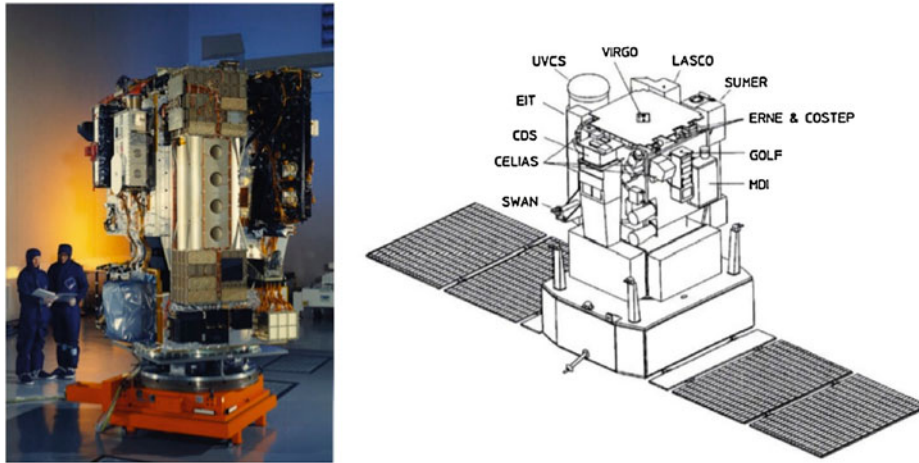


Fig. 6.2 The SOHO payload [46] was a complex suite of 12 instruments, including remote sensing imagers, spectrometers and in situ particles, plasma and fields experiments. The entire spacecraft

has a mass of 1,850 kg, is three-axis-stabilized and powered by solar panels delivering 1,150 W. The payload itself has a mass of 650 kg and a power consumption of 500 W in orbit. *Image ESA*

example and (as discussed in [Chap. 4](#)) telecommunications services from geostationary orbit, remote sensing of the Earth, Sun, or planet from a polar orbit, or orbit insertion around other bodies in support of deep space missions. Orbit selection feeds directly into the payload design process (and vice versa) in a number of ways. The orbit determines the range of environments to be encountered by the payload, affects the arrangement of the ground-satellite communications (see [Chap. 2](#)), specifies the observation windows whether for Earth, solar, or astronomical observations, and determines the data handling procedures of the payload through the available telemetry.

As introduced in [Chap. 4](#), orbits in common use for most space missions include Low Earth Orbits (LEO), Geostationary Earth Orbits (GEO), Sun-synchronous, polar, and a wide array of trajectories that are used to support missions that escape Earth orbit. This last category includes planetary and heliocentric missions that require flybys, orbit insertion, and sample returns, and can involve complicated trajectories, often encompassing multiple gravity assists. Such trajectories can generate severe constraints for the payload design, particularly in data handling and processing, telemetry, and environment control. LEO is commonly used for Earth observing, small astronomical, including solar, missions and, of course, Space Shuttle flights and the International Space Station [32]. GEO is typically used for weather monitoring, reconnaissance and telecommunications satellites. Sun-synchronous orbits are polar orbits with a rate of precession proportional to their orbital period around the Earth, as explained in [Chap. 4](#). This means that the satellite crosses a given location on the Earth at the same local time each revisit, which is useful for Earth science observations, weather monitoring, and military

reconnaissance. Polar orbits are geocentric orbits with high inclination angles (typically 75° – 105°) and are useful for a variety of purposes including mapping, telecommunications, Earth monitoring, and navigation. The final class of orbits covers a multitude of Earth escape trajectories that are predominantly used for scientific missions to distant objects within the solar system, including the Moon [47, 48], Mars [49] and the other planets (e.g. [9, 19], and the occasional asteroid [38] or comet [14].

6.1.8 Environment

As discussed in detail in [Chap. 3](#), the operational environment has a significant impact on the operation and safety of the payload and spacecraft. The interaction of the environment with the spacecraft structure and its impact on operations should be carefully considered in the design process. In addition to providing some significant challenges for contamination control (see later), the space environment can also lead to other operational difficulties, these will be summarized herein. The primary concerns are the effects of high-energy particle and photon radiation, spacecraft charging from the ambient plasma environment, atomic oxygen interactions in LEO, and dust impacts.

6.1.8.1 Radiation

The radiation environment is characterized primarily by the source of the radiation. All orbits are subject to electromagnetic radiation from the Sun, with ultraviolet and X-ray radiation being the most damaging because they can induce physical and chemical changes in exposed surfaces, act as ionizing agents affecting charging of the spacecraft, and

cause degradation of solar arrays. Ultraviolet radiation, particularly below 200 nm, is particularly damaging to solar cell, thermal control surfaces and optics coatings, causing significant degradation in the performance of these materials. Particle radiation derives from a number of different sources. For example, galactic cosmic rays, predominantly protons, reach the inner solar system with very high (up to TeV) energies and can pass through spacecraft leaving a swath of ionization damage in their wake. Solar flare and interplanetary-shock accelerated electrons, protons and ions have lower energies than the cosmic rays but their fluences are much higher and so they can have more frequent effects [50]. Finally, locally trapped particles around planets, like the van Allen radiation belts around the Earth, provide frequent periods of enhanced particle radiation for an orbiting spacecraft, e.g. passage through the South Atlantic Anomaly [51]. LEO spacecraft are shielded to some extent from cosmic and solar particles by the natural barrier to charged particles provided by the Earth's magnetic field. All of these sources of radiation can lead to problems with the spacecraft if not carefully prepared for. Effects include cumulative damage related to the total dose received, of single-event effects from individual ionizing events. Many sources of radiation doses are also time variable and sporadic, leading to widely varying conditions. Designing to the worst-case scenario is often not practical due to the costs involved and consequently, there have been a number of instances when large solar storms have led to the untimely termination of a satellite [52] as a result of a substantial and rapid enhancement in the local radiation environment.

6.1.8.2 Plasma

All spacecraft reside in an electrically neutral but highly ionized plasma environment, either due to a planetary magnetosphere or the solar wind (some spacecraft contribute to plasma environments through application of the maneuvering thrusters, venting of ionized gas, etc.). The interaction of the spacecraft with the rarefied plasma can result in the build-up of charge on the spacecraft structure that can then discharge and potentially damage electrical components of the payload. Spacecraft in low inclination geocentric orbits can spend one-third or more of their time within the shadow of the Earth. During this eclipse period, the spacecraft may negatively charge to tens of kilovolts, which can lead to a severe discharge when the spacecraft returns to daylight since this can result in positive charging creating a large potential drop across the spacecraft. Outside of eclipse, spacecraft in LEO tend to become positively charged because their typical velocities are greater than the ambient ion speed but lower than the ambient electron speeds. This creates a wake effect which then generates a differential charge across the structure that can generate strong discharges. In GEO, in addition to any eclipse

periods, internal dielectric charging can occur from the increased penetration of high-energy ambient electrons into the spacecraft that negatively charge insulating materials (e.g. printed circuit boards) or non-grounded conductors. Mitigating the effects of spacecraft charging can be accomplished by the careful selection of materials, the use of conductive reflective coatings for some optical components, particularly solar cell cover glass, to make the spacecraft as close to an equipotential as possible, and by defining appropriate ground paths in the system [53].

6.1.8.3 Atomic Oxygen

Atomic oxygen (ATOX) is a major environmental hazard for spacecraft in LEO. The flux of atomic oxygen within an altitude range of ~180–650 km above the Earth can lead to severe degradation of external surfaces, particular those made from polymers like Kapton®, other polyimide polymers, and Mylar®, and can, through scattering, lead to the erosion of internal surfaces. Measurements of the impact of ATOX on spacecraft surfaces on board a number of Space Shuttle Orbiter flights have shown that interactions with atomic oxygen can produce problematic changes in the mass and surface properties of a range of materials, primarily through the processes of erosion and oxidation. The motion of the spacecraft through the upper tenuous reaches of the Earth's atmosphere can lead to the oxygen atoms impacting with energies as high as 5 eV; both the velocity and the angle of the incident atoms are important. The severity of the accumulative effect depends upon many parameters, including orbit altitude, level of solar activity, and duration of mission [54]. Potential ATOX reactions include hydrogen abstraction, formation of radicals, oxidation particularly of metals, and oxygen inclusions into the C–H bonds of the polymers. The oxidation reactions, for example, can have a marked effect on the reflective properties of the surfaces including, most importantly, the optical surfaces of payload instruments. The suitability of certain materials for use in the ATOX-rich environment of LEO depends crucially on their erosion yield, i.e. the volume of polymer lost per incident atomic oxygen atom in cm^3/atom . Materials with high erosion yields, like most of the hydrocarbon organic materials, should be avoided or only used when protected by a low erosion yield material, e.g. fluoropolymers, or when they are not likely to come into direct contact with the ATOX environment. Mitigation of ATOX effects can be achieved by a careful choice of materials and coatings, and by carefully designing the payload so that sensitive surfaces do not come into direct contact with the atomic oxygen.

6.1.8.4 Micrometeoroids, Dust, and Space Debris

Micrometeoroids and dust orbit the Sun and their generally large velocities provide a direct hazard to the structure of

any space mission. These particles are much larger than individual atoms but small enough to be unobservable. Most micrometeoroids come from the plumes ablated from comets in the inner solar system, although a small fraction come from asteroids. Interplanetary dust, as its name implies, is scattered throughout the solar system, between the planets, and is mostly a hazard for planetary missions. Space debris is the collection of objects that remain from now-defunct missions, ejecta from control thrusters and rocket plumes, and particles of spacecraft material knocked off orbiting spacecraft via collisions with other objects. Space debris objects can be as large as a dead or dormant spacecraft, and it is estimated that several thousands of objects great than 10 cm across are tracked as orbiting debris.¹ Trackable objects can, in principle, be avoided by spacecraft maneuvering and objects smaller than about 1 mm across do not generally penetrate a spacecraft. This leaves the range 1 mm–10 cm as the critical size of particle in terms of spacecraft damage. In addition, a typical impact velocity in LEO is about 10 km/s for orbiting debris and about 20 km/s for micrometeoroids, fast enough for even a small particle to create significant damage. Near the Earth, the space debris environment presents the most significant hazard but in higher orbits, or for missions in hyperbolic orbits, the dust and micrometeoroid impacts provide the greatest danger. The small sizes and lack of knowledge of the size and velocity distributions of these particles require that calculations of the expected fluxes that can be encountered for a given orbit must be dealt with statistically. Protecting the payload from the collisions with micrometeoroid, dust, or space debris particles in the 0.1–10 cm size range presents a considerable challenge in the payload design. In a high flux environment the spacecraft and payload structure can suffer enhanced erosion, degradation of surfaces, and potential catastrophic loss of operational systems. Mitigation of these effects involves protecting the sensitive surfaces and components behind thicker structure or via multi-layer shielding, e.g. foils, which effectively shatter the incoming particles into a harmless spray on the inner walls of the shielding. This approach tends to be successful for particles up to about 1 cm [3]. More shielding would be required to protect the payload from the larger particles but often this is not a viable or attractive design option. In such cases, avoiding the main orbital traffic lanes or pointing sensitive surfaces away from the direction of travel are trades that should be considered as the design progresses.

¹ At the present time the orbital information for objects greater than 10 cm is publicly available. However, it is widely known that smaller objects are also being monitored but this information is classified and not in the public domain at the time of writing.

6.1.9 Contamination Control

Contamination control is crucial to the success of a given mission. This applies to on-ground manufacturing, integration, and testing as well as on-orbit instrument operations. Contamination can have several deleterious effects including degradation of optical and thermal control surfaces, degradation of the power system, electrical noise, and short-circuiting of electronics, and can lead to poorer precision on a variety of sensitive mechanisms. Very early in the design process, a contamination control plan should be developed and the design process should make every effort to minimize contamination at each stage. It is significantly cheaper to follow precise contamination control procedures on the ground than to correct for poor contamination control while in space.

The nature of the contamination depends crucially on the environment present at the various stages of the creation of the mission: laboratory/clean room, testing, launch, and space environments. On the ground, the largest sources of contamination are the people working on the payload. Human-borne contaminants can be introduced to the payload from clothing, breathing, sneezing, hair, skin, etc. Additional sources include the fallout from any nearby machining, particulate matter in the testing environments, and outgassing from the structural materials used. A number of contamination control documents exist at the various space agencies, both for airborne and surface particulates, organic and non-organic molecules, to define the requirements that must be met (e.g. [55–58]). These define the maximum acceptable levels of contaminant (particles per cubic meter of air), generally as a function of particle size. In space the primary contaminants depend mainly on the orbit selected. LEO missions suffer particularly from atomic oxygen, which can lead to strong oxidation of metal surfaces, react with the surfaces to produce volatiles, and chemically react with the surfaces to produce radiation, for example atmospheric glow as discussed in [Chap. 3](#) [59]. Other sources include micrometeoroids, space debris, and thruster plumes and outgassing from the spacecraft.

The effects of contamination can be kept within acceptable levels with the proper application of contamination control processes. Isolating sensitive surfaces from sources of contamination, using low outgassing materials and the vacuum-baking of components, and nitrogen-purging the instrument during assembly and ground operations can all serve to attain the required performance from the payload. Locating cables and connectors on the outside of the payload structure can facilitate contamination control during integration with the spacecraft. Use of witness samples in place of actual components can provide a direct means of understanding the level of contaminants. Other pro-active measures can be taken to protect sensitive

instruments during the outgassing phase when in space. For example, installing CCD decontamination heaters that activate shortly after launch can protect the CCD from thin layers of molecular contaminants, although larger-scale contamination can form small, localized structures that survive the bakeout process.

6.1.10 Verification

As the design and construction of the payload progresses, verifying that it meets its requirements is critical to producing a successful payload. The definition of the verification plan, with a carefully defined set of verification criteria for each payload element and a verification method (analysis, inspection, demonstration), should occur early in the design process. The actual verification process takes place later in the development of the payload. Verification can be costly, especially when requiring demonstration or testing, so the criteria to be verified must match well to the requirements and be precisely defined. It is very difficult to verify vaguely specified requirements, such as the detector will have a good signal-to-noise ratio. This frequently necessitates a translation of a given requirement into a verifiable statement: “the detector shall have a signal-to-noise ratio of at least 100:1 at a wavelength of 656 nm”. The verification plan will specify the components to be verified, the level of verification, whether it be at the component or systems level, the verification method to be used, e.g. demonstration, and the schedule for the verification.

The types of verification method used are: *test*, where a component of the payload is operated under specified conditions and compared with the requirement; *analysis*, where models or simulations of the component or system of components are used; *demonstration*, where the functional behavior of an operating component is shown to follow expectations; and *inspection*, where the designs and associated documentation are reviewed periodically to determine whether the requirements are being met. This is discussed further in [Chap. 7](#).

6.1.11 Trade Studies

Trade studies are an important component in the design of a payload. As the design matures from a set of concepts through downselect and final design, a number of critical decisions need to be made that not only respond to the choices available, the trades, but also to changing conditions, the trade-offs. It is important to make the key trades as early as possible in the design process because this leads to the biggest cost savings and the lowest risk. Some trades

and trade-offs are payload specific but trades at the mission level can have a severe impact on the payload design. Typical trades that impact the payload design include

- *Orbit selection*—Launch costs are significantly lower for LEO and the ground-to-space communications are less complex compared to other orbits, e.g. GEO or Sun-synchronous. However, for most commercial and military payloads the choice of orbit is fixed by the operational purpose of the mission. For scientific payloads, there may be some merit to an orbit trade-off even at the expense of some of the science return. For example, a high data requirement might benefit from the increased telemetry and ground-support from LEO but at the expense of continuous coverage of a given location on the Earth or of a specific astronomical object.
- *Spacecraft autonomy*—Autonomous operations can significantly reduce operations costs and complexity as well as optimize data collection and/or transmission. However, developing the ‘smart’ software for autonomous operations is expensive and a high level of autonomy reduces the flexibility of the operations and potentially increases risk.
- *Mission-specific flight software*—Developing new flight software is unavoidable for many one-of-a-kind payloads, and also for some repeat payloads. However, reusing software from previous missions when possible can provide a major cost savings and also provide a measure of reliability and lower risk. This comes again at the expense of flexibility.
- *Data management*—On-board data storage and processing can make for more efficient use of available telemetry compared to real-time data transmission, which provides information as needed but is operationally more complex. In many cases, especially those that collect large amounts of high-bit data (e.g. high resolution images), the available telemetry is the driving constraint that makes the choice of on-board data management significantly more attractive.
- *Technology trades*—A key consideration in the design process is the selection of the right technology for the objectives of the mission. The trade-offs here are typically in the category of performance versus cost, although for more advanced technologies at lower technology readiness level (TRL; see [Sect. 2.3.3](#) for a definition of the various levels) it may be more of a performance versus schedule issue. In addition, when comparing technologies the risk issues should also be borne in mind, particularly when pushing to those that have not been flight-tested. The range of applications where technology trades may be important can be quite large, including better thermal, electrical or mechanical control, better environmental control and mitigation, higher performance detectors, better structural rigidity, higher

accuracy guidance and control, use of previous flight spares, and so on.

- *Operational trades*—The primary goal of the payload is to produce the highest level of performance given the imposed constraints. Often, there is a degree of flexibility in the choice of the operational modes of the payload, where a bit of give and take can significantly improve the operational return. For example, spatial accuracy can be traded against temporal accuracy, or wavelength coverage for higher cadence, etc. These trade-offs are severely limited by the mission requirements and can lead to changes in other parts of the system that may be far from optimal, e.g. increased power requirement or higher telemetry. It is important to adhere as closely as possible to the operational requirements of the mission but to be open to some flexibility in the payload operation.
- *Risk versus return trades*—One of the most hotly debated areas when discussing trade studies for a mission is how much risk is worth how much gain. As the space industry has matured the assumption of risk has become anathema to most missions. Risk aversion is often the order of the day, and more often than not, the risk versus return trade discussions come down on the side of caution. However, the acceptance of a small or moderate increase in the risk to the mission may lead to significant enhancements to the operational or scientific return and it is often useful to consider all options during the early phases of the design. It is worth emphasizing again, that the earlier in the design process that the above trades are made, the less likely it is to run into scheduling and cost overruns and the more likely it is to have as near an optimum design as possible.

6.1.12 Conclusions

In summary, the basic design process involves a step-by-step approach through a number of well-defined tasks with some tasks being revisited as the design matures. The process starts with the generation of the project objectives and requirements, the identification of the various payload and spacecraft subsystems needed to meet these requirements, the development of implementation plans, and the analysis of any trade off. This leads to a conceptual design (see Fig. 6.1) that provides the basis for the detailed design of the payload. From the conceptual design and the trade-offs, analysis refinements to the program and mission requirements can be made and the payload specifications defined, in concert with an assessment of the costs and schedule. The detailed design then proceeds through the development of the selected concept and specific hardware/software specifications, the procurement of the necessary components and parts, and the manufacturing and assembly of the final payload. It is very important early in the detailed design

process to put together plans for verification and testing with a prescribed schedule and to perform these tasks at the appropriate points in the development. The final task, prior to launch and operation, is to verify, via testing or analysis, that all of the requirements are being met by the final payload. In particular, a sensitivity to the environment in which the payload will operate for the bulk of its functional life is crucial. The radiation, both particle and electromagnetic, dust and magnetic conditions that surround a spacecraft tend to vary with location in the orbit, with time as solar conditions change, and with distance from the Earth: typically a payload will have to operate safely in all of these environments. The verification process must be performed in conditions that are as close to nominal as possible, while measures should be implemented to mitigate the expected worst-case scenarios, e.g. passage through the South Atlantic Anomaly [51].

6.2 Mission Resources

After the initial experiment design process has developed a set of science requirements and corresponding set of measurement requirements, it is important to determine what is required from the spacecraft system in order to support the proposed payload. These factor into the development of the overall mission requirements, or as part of an iterative process to optimize the overall design as discussed in the previous section. The resources needed to accommodate the payload can be categorized as those resources needed directly by the payload, mass and power for example, and the additional resources required in other spacecraft systems to support the payload, e.g. additional structural mass, spacecraft pointing precision or telecom antenna size, or programmatic resource constraints, such as redundancy, risk, technology readiness or margin. The resources required by the payload are ultimately captured in the interface specifications that describe all the aspects of the connection between the payload and spacecraft, including mechanical and thermal interfaces, power requirements with associated voltages, and telemetry and commanding interfaces. Often there will be requirements on electromagnetic cleanliness and acceptable interference levels levied by the payload on the system as a whole (including components of the payload on each other) and levied by the system on the payload (see contamination discussion in Sect. 6.1). Again, these are captured in the interface documents.

The most obvious payload needs are mass, power and telemetry bandwidth and are usually quite straightforward to determine, although there are subtleties such as cable harness mass, detailed estimates of which require knowledge of the spacecraft bus configuration. If mass is a major driver (which is often the case) a number of design choices

can be made to reduce the overall payload mass. If the payload consists of multiple instruments, then it may be possible to share resources between instruments. A common data processing unit or housing electronics cards from several instruments in one chassis can significantly reduce mass in some cases. These approaches may have impacts in systems engineering and integration and testing, but don't usually drive costs. More aggressive approaches include the use of magnesium alloys instead of aluminum for instrument structures, or using composites for optical benches (this is sometimes done to improve rigidity and reduce thermal expansion in optical systems). These lower mass structural elements can significantly reduce the mass constraints while also improving performance but are expensive and, in some cases, may require custom builds, all of which increase cost and may impact the schedule.

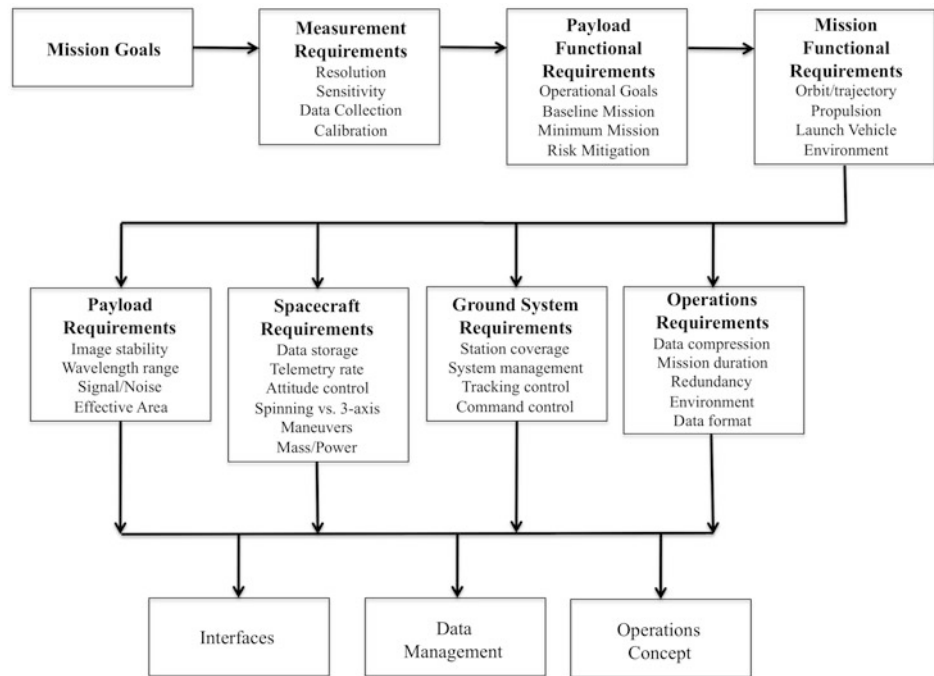
Telemetry bandwidth can be a key driver of mission complexity. Both the bandwidth needed to return the payload data to the ground and the bandwidth needed to move the data through the spacecraft bus are important, and can be significantly different, depending upon on-board processing and data compression. For high bandwidth missions, telemetry bandwidth can form part of a complex trade between the telecom system (antenna size, antenna pointing control, RF power), on-board computing capability, on-board data storage, payload mass, and power. A significant driver is often the quality, continuity and timeliness of the returned data: what data loss is allowable during transmission, what observing duty cycle is required, and how quickly does the data need to be returned to Earth? If data can be stored for long periods on the spacecraft before being returned to ground, then the telecom system and operations concept can often be simplified: if data is required soon after collection this can be a significant driver on the complexity and cost of a mission. Data management for the payload includes the selection of telemetry bandwidth, the use of ground-based antennas for data collection and the on-board storage and compression of data, and involves trades in all these areas (see the following section and [Chap. 2](#)).

Instruments often require control and knowledge of their orientation in terms of the direction they need to view, the required level of control and knowledge of that direction, and the required field of view. Concerns that need to be identified in concert with determining pointing requirements are the potential for mutual interference, for example by obscuration of fields of view in optical systems, or side-lobe interference in RF systems, and required observation strategies. The latter can again lead to a complex set of trades: if a telescope system needs to collect data for a few hours a day, then the spacecraft can be reoriented to allow high bandwidth data return, if continuous data collection is required, then the telecom system must be more complex, using, for example, a gimbaled antenna.

It is rare that a payload system does not require some control of its operating temperature. This could be to keep electronics within tens of degrees of room temperature, to control an optical bench to within 0.1 °C, or to maintain a focal plane at a temperature of −50 °C. In each case however, interactions with the spacecraft play a significant role in determining how difficult the requirements are to achieve. Thermal control becomes particularly important when part of an instrument requires significant cooling. This occurs most often with imaging systems, when visible focal planes need to be cooled to reduce dark current, or infrared focal planes and optical systems need cooling to prevent thermal noise swamping the desired system. Passive cooling can sometimes be used for cooling visible light CCDs, with a radiator on the outside of the spacecraft coupled to the CCD via a conducting strap. This is the simplest approach, but does impose an additional constraint on the spacecraft orientation, to avoid solar illumination of the radiator. Maintaining temperatures below about 100 K requires an active cooling system: this could be as simple as a thermoelectric (Peltier) cooler [60], or for even lower temperatures a cryostat (essentially a large thermos flask) or an active cryocooler. Missions have been launched that used cryostats cooled with solid hydrogen, e.g. the Wide-field Infrared Survey Explorer (WISE) [61], or liquid helium in the case of the Spitzer mission [62]. While cryostats can be simpler to implement than active cryocoolers, they do impose a finite lifetime on the observing system (see discussion in advanced technology cryocoolers in [Sect. 6.5](#)).

The control and commanding of payload elements can also drive the need for resources, both on the spacecraft and on the ground. Control can be required for simple operations such as changing an instrument mode, with no interaction with the rest of the payload, or for complex activities such as reorienting the spacecraft to capture images. Commanding a payload to carry out these operations can be done via a stored sequence of commands on the spacecraft, which can be periodically uplinked, or in 'real-time' by sending commands from the ground when needed. Both approaches have their advantages—for example, the latter gives more flexibility to respond to events, while the former is less impacted by the availability of a telemetry link. The computing power needed to operate an instrument, or entire payload, both in processing and executing commands and controlling the basic operation of the instrument can be a significant driver on resource requirements—simple instruments can be controlled as state machines, with firmware executed by field programmable gate arrays (FPGA), while more complex operations may require powerful CPUs with complex flight software. A complex hardware and/or software design can significantly increase the overall cost of the payload, particularly given the need for more sophisticated test and verification efforts.

Fig. 6.3 Payload design drivers and interconnections



However, such costs may be justifiable if they result in a significant savings in the operation costs. An aspect of payload control that is receiving increasing interest is the use of on-board autonomy to control aspects of payload operation or data manipulation. A good example of this type of capability is sometimes seen in payloads that study space plasmas in situ. NASA's Time History of Events and Macroscale Interactions during Substorms (THEMIS) mission [9] used a 'burst mode' capability to change the operating mode of multiple instruments in response to triggers derived from data collected by each spacecraft (THEMIS had five separate spacecraft). This allowed the spacecraft to capture high rate data around interesting events in the magnetosphere, without overburdening the telemetry link. This type of autonomy is key to capturing sporadic events that cannot be predicted and are too brief to be captured by a response from Earth.

Finally, programmatic drivers on a mission can have a significant effect on the resources required by the payload. In particular, the need for redundancy can drive mass and complexity as can the imposition of large margins. The latter can be mitigated somewhat by using high-heritage instruments, where the potential for growth is more predictable, but often payloads require capabilities that go beyond the current state of the art (larger format CCDs, for example). Balancing the desire for enhanced performance against the cost and complexity of a payload is a key part of the payload design process.

The interplay between the mission resources and the ultimate capability of the payload is important in the design of a successful mission. Limited resources can severely

impact the performance of the payload while a poorly designed payload operations plan can severely tax the available resources. While there may be some flexibility in the mission to accommodate additional resources, this is rare and, in any event, optimal performance is best achieved when the resources and payload capability are closely matched early in the design.

6.3 Design Drivers

The opening sections of this chapter addressed the basics of payload design featuring an overview discussion of the various high-level factors that impact the design of a space payload, whether it be for scientific, commercial, or other purposes. In this section the emphasis is on the particular factors that drive the design. In other words, while addressing all of the factors discussed previously is necessary in order to generate an optimized design for the task at hand within the various budget and scheduling constraints, there are specific factors that have little or no flexibility if the overall mission goals are to be achieved. These primarily come down to how the various top-level requirements are defined (see Fig. 6.3).

6.3.1 Performance Requirements

The most important attribute of a space payload is its performance: will it meet or exceed the required level of operation throughout the lifetime of the mission? Designing

and sizing the payload to satisfy the performance requirements necessitates a clear specification of the objectives of the mission, the flowdown into the payload characteristics, and the implementation of these characteristics into the final design of the payload. This flowdown is described in terms of different levels of requirements. For the payload, level 1 requirements capture the high-level objectives of the payload, and level 2 requirements capture the required instrument. The transition from a conceptual design to the detailed design was discussed earlier, following the considerations of cost, schedule, the various trade-offs, and integration. Here the focus is on the specifics of the payload operational design.

The first stage of any mission is to specify as completely as possible the operational goals of the payload. This is often articulated as a set of requirements that flow down from the highest level mission requirements first to a set of measurement or performance requirements and then down to a set of instrument functional requirements. For scientific missions this leads to a set of scientific instruments designed to answer a specific top-level set of science questions. An overall strawman payload design may be produced by a science definition team prior to a competitive selection process and is then made more rigorous during the conceptual design stage (or Phase A). Commercial and military payloads have a very different flight rationale, and as such the goals of the mission tend to be more focused on the direct operation than the broad top-level goals of a science mission. However, the payload design must respond to these high-level requirements, and the expected payload performance must be traceable back to these requirements. The commercial payload, then, is designed to meet a specific operational goal, e.g. to provide continuous telecommunication connections within a specific geographic region, with a certain precision, accuracy, and sustainability under a very well defined set of constraints, such as minimum signal strength and data rate over a given bandwidth for a minimum number of simultaneous users. Similarly, military payloads typically focus on safety, reliability, ease of operation, production and testing, and, of course, affordability. As such, military, and for that matter most commercial, payloads pay more consideration to redundancy and have longer design lives than most scientific missions. Moreover, military and surveillance payloads frequently have to be responsive to events that happen on timescales shorter than a typical design-to-build duration, and therefore having a standardized system with some built-in flexibility, i.e. a spacecraft that can accommodate a range of payload sizes, is desirable but has a significant impact on the design of such payloads.

Most missions are based on a response to an Announcement of Opportunity (AO) or Request for Proposals (RFP) and so can be constrained in advance by the

available funding, limited range of spacecraft and launch vehicle options, and, often, on a preset scientific focus defined in advance by a specially commissioned study panel, specific program goals, direct commercial or military need, roadmap exercise, or international agreement.

Whatever defines the constraints, the primary purpose of the early design stage of the payload is to define the specific instrument performance requirements needed to meet the stated goals and identify what requirements the payload places on the spacecraft and mission as a whole (power and telemetry requirements, for example). This will typically require a number of compromises (trade-offs) and refinements to be made as the operational goals are translated into a working design, and often trade studies that compare different measurement approaches or mission architectures. This is where there is a significant difference between scientific missions and most commercial or military missions. Frequently, the predefined goals of a scientific mission are broad and attack a particular scientific phenomenon and class of observation giving significant leeway in nature and performance of the various conceptual payloads proposed to meet these goals. Entire payloads, or individual instruments are usually selected competitively via a proposal process, with the science and science implementation reviewing the resulting proposals and selecting the best conceptual design that meets the performance requirements and seems likely to achieve the operational goals within the stated constraints. This conceptual design forms the basis of what will develop into the Baseline Mission, while a subset of this design, one that meets the minimum performance required to meet the operational goals, is generally referred to as the Minimum or Threshold Mission. The Baseline Mission is designed to accomplish the entire set of identified mission objectives while the Threshold Mission provides a measure of the minimum expected performance of the payload and is the worst-case fallback position should the development of the baseline payload fail to meet the performance, schedule, or cost targets. Any performance below the Threshold Mission level is deemed not to be justifiable at the proposed cost. The Minimum Mission should be sufficiently different from the Baseline Mission in terms of its key capabilities and ability to meet the mission objectives. Simple de-scoping of the instruments that don't result in cost or schedule savings, or shortening of the operational duration of the mission to make up for cost or schedule overruns are generally not considered to differentiate the two concepts. Should the development of the payload run into problems, a prioritized plan to reduce the capability, with respect to the performance requirements, is required during the concept study (Phase A) period. This plan should consider key risk drivers, approaches for their mitigation and the full range of possible de-scoping options, with triggers for each such option.

6.3.2 Measurement Requirements

The ability of the payload to perform to specifications and to meet the overall mission requirements depends crucially on the ability of the payload instrumentation to make the requisite set of measurements with sufficient precision. The measurement requirements are the major drivers on the design of the payload because the operability of the instruments requires a range of resources and accommodations. This is particularly true for complex payloads with multiple instruments. Moreover, the ability of the instruments to meet the measurement requirements has to be verified during the payload development and they must be calibrated both during the design and in space. This flow of requirements is captured in the traceability matrix, which relates a specific measurement objective downwards to a specific function of a given instrument and upwards to the relevant mission objective. The payload design has to accommodate the operation and calibration of the instruments and the validation of the measurements they make.

The primary issue that drives the design is therefore the requirement that the instrument(s) meet or exceed in the dynamic range, accuracy, sensitivity (signal-to-noise), resolution (spatial, spectral/energy, temporal), and throughput thresholds defined by the overall mission objectives. Furthermore, the data management, payload operability, measurement calibration, and cadence of measurements all factor into the design, the interface with the spacecraft, and spacecraft operations.

As discussed in the performance requirements section above, scientific, commercial, and military payloads often emphasize different driving factors to optimize in the specific set of measurements or observations required. All of the above parameters influence, and are influenced by, spacecraft operations (e.g. 3-axis stabilized or not) and mission specifications (e.g. choice of orbit), further emphasizing how an optimal design is one where all of the system elements are considered together.

It is not always guaranteed that all of the measurement requirements can be met in a single operational mode of an instrument. The signal-to-noise ratio of a given measurement is critical regardless of the required performance. Depending on the operational goals of the mission, this requirement can be very strict or relatively lax but to make meaningful use of the data collected the signal always has to be distinguished from the noise. In an operational environment where photons (or particles) are limited, e.g. astrophysical observations or low-contrast reconnaissance observations, there is a direct competition between the various temporal, spatial, and spectral resolutions and the required sensitivity. Consequently, the throughput becomes a controlling factor. The design challenge is to maximize the throughput subject to the various constraints on the

mission. In addition to the usual constraints of cost, schedule, mass, power, volume, etc., constraints which directly affect the measurements and that, in most cases, tend to lower the throughput must also be included. Such constraints include the trades between field-of-view and spatial resolution, aperture size and scattered light management, wavelength coverage and signal discrimination. There are a number of ways to meet these challenges and these should be reflected in the range of conceptual designs that essentially initiate the payload design process (see Fig. 6.1). For instance, a simple solution to many problems would be to build a bigger, more complex, and more capable payload, assuming there are no restrictions imposed on cost, schedule, launch vehicle, etc. However, this is typically not a viable option. More feasible approaches include

- Carefully planned selection of operation modes where photons are shared differentially between the various observational modes (e.g. spectral resolution at the expense of spatial resolution, signal to noise at the expense of temporal resolution).
- Autonomous capability to respond to changing conditions in the quantities being measured (e.g. changing exposure time, localizing or widening field of view, changing magnification, collecting bursts of higher cadence data).
- Simplifying the payload to focus on providing a set of routine measurements where the design compromises, between the different instruments or instrument modes, and meets the operational goals but does not push the limits (e.g. fixed sequence of observations with each observation having a fixed set of parameters).

A more complex operational approach often produces a more costly and risky design, along with more expensive day-to-day operations. Simple, routine, synoptic observations often involve the least risk but at the expense of tailored operation and flexibility. The key factor is to weigh the scientific or operational return against the increased complexity and risk.

The ability to satisfy the measurement requirements is not reliant on the instrument operations alone. Restrictions imposed by the data management system must also be considered. Data management includes the capability of the payload or spacecraft to collect, store, and transmit data, and this is intimately tied to the measurement requirements on the one hand, and the ability to transmit the data to the ground on the other. The expected data volume to be handled, how this data is managed on board, and how it is transmitted to the ground, all factor into the allowable throughput. If real-time data is required then the spacecraft must be in contact with a ground station and the data volume is limited by the telemetry rate, which is, in turn, determined by the communications system, the orbit, and the data compression approach used. If downlink time is limited, say to several 10 min passes per day, then only a

limited amount of real-time measurements can be made and operators must then rely more heavily upon on-board data storage and subsequent transmission. On-board data storage does not, unfortunately, remove the telemetry limits, but it does allow the operators to utilize on-board processing to pre-select what data gets downloaded: data above a certain count rate threshold; data that meets some predetermined behavioral or fidelity criteria, for example. On-board storage can also provide operational simplifications by decoupling payload observations from data downlink by using only a high-power transmitter intermittently, or pointing a high-gain antenna at the ground station when high-rate data are being returned. In designing the payload, consideration must be given to the data handling, in particular how much and what kind of data is to be stored for subsequent downlinking, as this will feed directly into the operation of the payload and will affect the instruments' ability to meet the measurement requirements.

In order to verify and validate whether the payload is meeting the measurement requirements while in space, it is important to take and transmit instrument, payload, and spacecraft health and safety data (e.g. operational temperatures, battery charge, communications status). In addition, different mission types place different levels of importance on the fidelity of the data and are more or less tolerant of data transmission errors. Commercial payloads place a lot of emphasis on maximizing the data rate and minimizing data errors. The requirement of a high data rate impacts the on-board communications system by requiring either large antennas and/or higher transmitter power. The design trade here depends on whether the spacecraft can accommodate the necessary increase in power and/or antenna size. Accommodating a high data rate is also important for high-throughput scientific payloads, especially those without a continuous dedicated downlink (i.e. a less than 100 % duty cycle). The bit error rate denotes the probability of a bit error occurring in the data transmission either within the spacecraft or between the spacecraft and the ground. The higher the bit error rate the more likely it is that data will be irrecoverably lost, although error correcting codes in packetized telemetry will recover some lost bits. In addition, there are a number of encoding schemes that are designed to enhance error resistance, e.g. low-density-parity-check (LDPC) codes [63]. For most scientific payloads the occasional loss of data is frustrating but manageable. This may not be true for commercial or military payloads where such a loss could lead to significant financial impact or critical loss of information.

One last data management issue of great relevance to meeting the measurement requirements is the use of data compression and whether to allow for lossy or lossless compression. Data compression significantly increases the amount of data that can be stored and transmitted, and by

that token is a good thing. Lossless data can exceed factors of 3:1 or 4:1 (with some wavelet-based algorithms claiming much higher lossless compression [64] and so potentially a great boon to data intensive payloads). In some cases lossy compression may be tolerated although as the name implies, this will lead to the loss of information and thus should be used with care. Some data are more amenable to compression than others, and a careful analysis should be carried out in advance to determine the worth of applying data compression because it comes with some time, cost, and a little additional complexity to the spacecraft operations.

The operability of the payload can have a significant impact on the payload design. Complex payloads require more complex operational procedures that, in turn, require sophisticated software and additional mechanisms with associated increase in the on-board computing, mass, and power needs. This all adds to the cost of the mission by increasing the verification and validation load, and raises the risk. Conversely, the push for complex payloads to improve the flexibility of the operations will increase the data load and lead to compromises or trades being made which may impact the ability to meet the measurement requirements. Operating the payload, then, becomes a driver of the design process.

Payload operations can be classified in a number of ways but the most common are: *survey payloads*, *event-driven payloads*, and *adaptive payloads* [65]. As their names suggest, *survey payloads* have relatively simple operations and at their most basic they rely on repeating the same suite of observations or operations each orbit, *event-driven payloads* are designed to make a series of operations at specific points in their orbit or at given times of year, and so their operations tend to be more complex, often requiring accurate time management and variable spacecraft pointing, *adaptive payloads* are the most complex as they are required to respond to changing circumstances that either naturally occur or are commanded as part of the mission plan.

Each operations concept must include planning, scheduling, commanding, data management, spacecraft support (e.g. roll maneuver), and response to physical conditions or events. The more complex the planned operations the more detailed the operations concept has to be. In addition, all of the operations must be performed while meeting the various performance and measurement requirements. In designing the payload operability, one of the biggest decisions is whether to adopt autonomous or commanded operations (sometimes a mix). As discussed, some of these decisions are influenced by the nature of the measurements to be made, their purpose (e.g. real-time response), and the volume to be transmitted. Additional factors include spacecraft capability, telemetry, orbit, data fidelity, etc. These considerations also require relating the payload design to the ground support (see [Chap. 20](#)).

Measurements are only as good as the level of trust placed in them, and the measurement requirements provide the necessary definitions to guarantee sufficient fidelity to meet the operational goals of the mission. A high signal-to-noise ratio and a low data error rate have already been identified as possible drivers of the payload design. However, adequate calibration of the signal must be provided to have confidence in the measurements being made. Calibration of spatial resolution, wavelength, pointing accuracy, spacecraft maneuvers, together with assessment of dark current and flat field levels (e.g. for CCD imagers) are all necessary and should be performed at regular intervals during the design phase and while in space.

Ground calibration is often performed at the subsystem or individual instrument level prior to integration of the payload with the spacecraft. In addition, in space calibrations are important and can utilize calibration sources included specifically for this purpose or regular observations of natural calibration sources. For instruments where neither of these is possible, external calibrations can be developed. For example, the Extreme Ultraviolet Imaging Telescope (EIT) on board the Solar and Heliospheric Observatory (SOHO) spacecraft has utilized specially designed sounding rocket calibration flights, the EIT Calroc [66], to perform benchmark calibrations of the EIT detectors. The rockets incorporate a copy of the EIT space optics and collect sufficient data during their operation to calibrate against observations by the contemporaneous and identically designed SOHO EIT. This has allowed for a number of calibration updates over the course of the mission since its launch in 1995.

Calibration is an important component of the instrument, payload, and spacecraft design. The level, number and complexity of the various calibrations can interrupt the regular operations of the payload, add to the complexity of the design, increase the data volume, and add cost, but they are essential.

6.3.3 Spacecraft Requirements

The payload is, of course, part of a larger system, and is reliant on other parts of this system for its successful operation. In addition to supporting the payload structure, providing power, removing heat, and supplying a host of other resources to keep the payload operating, the spacecraft also provides crucial capability to enable the payload to meet its operational goals and to satisfy its measurement and performance requirements.

The spacecraft generally provides the command, control, and data handling that is critical for the operation of the payload and the recovery of the data that it collects. This is discussed in detail in the previous section and while

important, it does not directly contribute to the taking of measurements. Instead, here, the focus is on spacecraft maneuverability and attitude control, which are critically related to the ability of the payload to make the measurements to meet its operational goals.

Spacecraft attitude control, the precision to which it can hold a given attitude, and the accuracy of the knowledge of its attitude, all directly affect observations made by the payload, particularly for payloads focused on target acquisition (e.g. imaging the Earth or an astrophysical object, and fixed station-keeping for continuous communications or surveillance). The ability of a spacecraft to provide a given level of pointing control and the stability to maintain a given position can drive the design of the payload control mechanisms, either to provide a finer tuning of the pointing or to control the effects of spacecraft jitter. In remote sensing payloads, there is often an important trade to be made between the capability of the spacecraft to control the pointing stability and the inclusion of an image stabilization capability as part of the payload. Depending on the required pointing stability, a substantial reduction in cost and complexity can be realized by having the payload carry some of the stabilization requirements.

Attitude determination and control can be accomplished in a number of ways, each of which affects the payload design differently. Conversely, a specific measurement requirement for the payload may affect the choice of attitude control system on the spacecraft. As will be discussed in Chap. 12, the most common approaches adopted for spacecraft attitude control are to spin-stabilize the entire spacecraft or to have specific 3-axis spacecraft control using a combination of gyroscopes and either reaction wheels or thrusters. Other approaches may use gravity gradient stabilization in Earth orbit or magnetic torquers that react against the Earth's magnetic field to apply a torque to the spacecraft. Spin stabilization and gravity gradient control can be largely passive, with overall spacecraft attitude controlled 'open loop' (spin period can be sensed with Sun-sensors and controlled via thrusters). Three-axis stabilization requires a control loop, with the control signal coming from a star tracker or camera, or via inertial sensors for less precise control requirements.

The spacecraft attitude control system (ACS) points the payload at the target within the required tolerances (see the traceability matrix discussion in Sect. 6.1). Finer control within the payload itself relies on additional instruments such as guide telescopes that provide fine pointing angular measurements (using known sources) and feed error signals to the ACS for spacecraft-level refinement of the pointing, as well as to an image stabilization system, if present. The image stabilization system is used to correct for spacecraft jitter that can smear high spatial resolution images, particularly during long exposure observations. Often payload-

based image stabilization systems operate with a different frequency bandwidth and correct for jitter at high frequencies that lie outside the bandwidth of the spacecraft ACS. For solar observation missions, for example, either a guide telescope can use the solar limb as the reference source or post-processing using feature-correlation tracking can be used to refine the pointing knowledge and remove spacecraft jitter.

In addition to accurate control and knowledge of the spacecraft attitude, the payload design may call for specific maneuvers such as spacecraft roll, large-scale pointing shifts, say to a change in target or center of field-of-view, or station-keeping. This adds complexity to the operations of the spacecraft and needs to be designed in conjunction with other mission considerations, such as preserving ground communications, maintaining the power system, thermal environment, and so forth. For complex payloads with conflicting pointing requirements, the payload may contain a scan platform to independently point one or more instruments.

6.3.4 Mission Constraints

While the focus in this chapter is on the payload design, with the overall space system design being discussed later in the handbook, the interdependency of the payload and mission can have significant consequences for the payload. Some mission drivers were discussed in the previous section. Here, some of the more important aspects of this interdependency provide the focus for the discussion.

Various places throughout this chapter have discussed the interfaces between the payload and the spacecraft and considered how some of the mission criteria impact the design. However, some constraints imposed by the mission on the payload need further attention.

The most obvious area where the mission and payload intersect is in the choice of the orbit and/or trajectory. Many factors that contribute to the orbit decision do not necessarily take the performance or measurement requirements of the payload into consideration, e.g. choice of launch vehicle and launch date, maneuverability of spacecraft, ground-station support, health and safety considerations (especially for human space flight), etc., and the payload typically tries to optimize its operations to accommodate these restrictions. However, in many cases the orbit is defined by the payload requirements, e.g. polar observations/communications, Sun-synchronous or geostationary operations, specific flight trajectory, etc., and then the design of payload goes hand-in-hand with the design of the orbit-related mission criteria.

One of the main issues associated with a specific orbit, in addition to satisfying the primary purpose of the mission, is the impact on the operations of the payload. For example,

LEO limits up/downlink connections to brief and separated pass times, leading to more complex operations. However, LEO is easier, and so cheaper, to attain, allowing for larger payload masses to be considered. The radiation environment of LEO (outside of South Atlantic Anomaly passage) is more benign than other, higher, orbital choices, but the space debris and ATOX environments are more hazardous. The specific issues associated with each orbit/trajectory are manageable, but not without due consideration and not without being accommodated in the payload design. Higher shielding against radiation for geostationary orbits, scheduling more ground station passes for LEO, and better station-keeping control, can all be factored into the payload design and operations to maintain required performance levels within the given orbit.

The payload sensitivities also play a role in the spacecraft operations and these also vary for specific orbits. In particular, for LEO, passage through the South Atlantic Anomaly when the radiation environment is particularly intense, requires, in some cases a different operational mode for the payload, requiring lower resolution observations, application of a filter or shutter, or complete shutdown of the operations (e.g. the Hubble Space Telescope). In addition, operating during eclipse periods may require a modified approach. Conversely, avoiding bright objects, like the Sun, which may damage sensitive detectors designed for less intense illumination, becomes a factor that also varies with orbit selection.

Interplanetary trajectories pose their own problems depending on whether the intention is to orbit a distant object, flyby with imaging and environment measurements, sample and return, or landings. While consideration of these is predominantly in the purview of the mission operations, they can significantly influence the payload in many ways: data storage and transmission needs specific tailoring for the low telemetry resulting from the large distances involved; autonomous control capability greatly simplifies the operability of the payload; transition from *en route* operations to on-site operations, especially for very long flight times, increases risk.

Another prominent mission constraint that is crucial to the payload is the health and safety status of the mission as a whole. It is obviously important to monitor the general well-being of all systems and their interactions. The payload needs to be able to respond (either autonomously or by command) to a change in the spacecraft and/or mission conditions. This is generally known as fault management and is typically designed to be managed by the payload and spacecraft through procedures defined in the Interface Control Document (ICD). The occurrence of an anomaly in the mission needs to have an appropriate response in the payload. The most common approach is to place the payload (and/or spacecraft) into a safehold state, also called safe-mode. When

in safe-mode, all unnecessary mechanisms and autonomous operations are shut down except for the communication channels with ground support. Once the anomaly or problem has been diagnosed, its effect assessed, and, if need be, corrected, the payload can be ‘recovered’. Full and safe recovery from a safehold state is a further requirement imposed upon the payload. A payload with a well-designed automatic safe-mode entry and recovery process can reduce the amount of health and safety monitoring required, thereby reducing the time and cost of real-time operations.

In addition, building in a range of positive performance and operational margins into the payload can significantly reduce the level of payload monitoring. Having more power capability in the spacecraft than required by the payload can mitigate problems with the power generation (damage to solar panels for instance). Having a larger radiator, to remove excess thermal loads, than required for expected operational heat build-up can avoid detector damage. And so on. Typical margins for most spacecraft systems and subsystems are targeted to be $\sim 30\%$.

6.4 Design Constraints

While the general flow of the design process is to develop requirements from the mission objectives (science goals, for example) and determine payload operating requirements and system requirements from these, payload requirements are rarely formulated without some a priori understanding of the constraints that a mission will face. In fact, such constraints are often key drivers on the design of a payload and the scope of the mission objectives. There are many constraints that are imposed on a given mission which vary with the particular goal of the mission. Scientific, military, and commercial payloads suffer from disparate conditions imposed by the mission but some are common to all and can be illustrated effectively by considering the details of a specific class of missions, for example NASA scientific missions.

For NASA missions, the most tightly constrained payloads are typically those for competitive missions (e.g. the Explorer, Discovery or New Frontiers programs), where a single principal investigator leads an entire mission investigation. Many of these constraints are simple restrictions on payload resources, such as available bus voltages, but others can form a complex trade space, for example mass could be constrained by a combination of launch vehicle capability, mission destination, required power (translating to solar array mass), communication requirements (which drive power requirements and are driven by spacecraft-Earth distance), etc. An example of how such a constraint is communicated to a mission is by a limited choice of launch vehicle capabilities, as in the 2011 NASA Explorer Announcement of Opportunity (AO).

Table 6.4 shows a subset of the launch constraints for this particular AO. In addition to maximum lift performance as a function of chosen orbit, each potential launch vehicle may have different payload fairing volumes and could provide different launch vibration environments. As can be seen from the table, there is a large range of available maximum performance, with the actual available performance being potentially much less than this for any given orbit, so the choice of orbit in such a constrained mission could significantly affect the available payload mass by many hundreds of kilograms. Thus, the science requirements that drive orbit choice can potentially severely limit payload options.

In determining the constraints that will be encountered by a particular payload, it is instructive to start by determining the requirements that the payload will levy on the overall system by determining the payload mass, power, telemetry bandwidth, on-board computing, thermal requirements, etc. Within the payload, resources can sometimes be shared to minimize the system impact. Examples are shared low voltage power supplies (a single power converter to convert and condition bus voltages to those required for digital circuits, or housing multiple electronics cards in a single chassis) although careful engineering analysis is required to make sure such trades are feasible and don’t introduce unwanted problems such as noise in electronic systems.

It is important to note that the addition of a payload, or payload elements, will consume system resources above these estimates. For example, increases in payload mass can result in greater structural and propulsion system mass, and increased power requirements translate into heavier power systems. Subsequent phases of payload optimization take these effects into consideration.

Depending on mission type, these design trades can take place at different times. For payloads where individual instruments are solicited separately, payload resource optimization comes after instrument selection, but when the payload is solicited as a package, much optimization can take place before selection. Similarly, when an entire mission is competitively selected, significant design optimization can take place before the mission is selected for funding. Experience with NASA competitive missions suggests that this latter approach is a factor in reducing overall mission cost growth.

6.5 Impact of Future Technology Developments

The mechanics of payload design have been honed and refined over the years so that documents such as this can provide most of the rudiments that apply to most payloads,

Table 6.4 Maximum launch vehicle capabilities provided with the NASA Explorer 2011 announcement of opportunity

Launch site	Assumed inclinations	Altitude range (km)	Maximum performance (kg)
CCAFS	28.5°–51.6°	200–2,000	1,585
KLC	70°–90°, SunSynch	200–2,000	1,640
RTS	0°–90°, SunSynch	200–2,000	855
VAFB	60°–90°, SunSynch	200–2,000	1,390
WFF	45°	200–2,000	1,435

This table shows a subset of the available capability

with much of what's missing being applicable to payloads with added complexity. As the technology of space exploration advances and as we seek newer means of observing in space, the payload design process will have to adapt to meet new challenges. In some cases, for example the low power, lightweight, electronics provided by nano-technology, the design process may be simpler and more effective, whereas in others, such as the need for cryogenic detectors, the design may become significantly more complex. In this section, some future technology needs and their impact on the payload are highlighted. This is by no means an exhaustive list nor is there space to be comprehensive but there is hopefully enough to give a flavor of how important innovation is in designing space systems.

6.5.1 Power Technology

For many payloads, one of the major system drivers is the power that they need. Issues of sizing the solar arrays, sizing the batteries, sharing the available power between all of the systems demanding it, and maintaining these resources throughout the mission lifetime all have to be addressed without exceeding the resources available to the mission. To-date most of the effort in making more effective use of on-board power has centered on the development of low power mechanisms and instruments, rather than on the power generation. A number of recent advances are pointing the way to more efficient power generation and use which, in turn, may facilitate the use of more complex payloads.

Two ways in which the power management of a space mission can be significantly improved is to provide more power per unit area of solar panel and to provide more efficient batteries for storage of the incident solar energy. Recent breakthroughs in the field of nano-technology are getting closer to providing systems of the scale required for use in space payloads.

Nano-technology research has led to the development of solar panels in which nano-designed particles, known as quantum dots, replace the silicon wafers in the panel semiconductor. Quantum dots are efficient at capturing photons and converting their energy to electrical current,

with tetrapod geometries proving to be amongst the most efficient. Tetrapods have the unique geometry of always having an axis pointing outwards, increasing the light absorption in the solar panel. Tetrapod-infused solar panels can potentially increase the energy conversion efficiency to as high as 70 % compared to the current value of around 20 % [67]. Currently, the maximum energy conversion efficiencies of tetrapods are around 10 %, significantly lower than their silicon counterparts. This is due, in part, to the difficulty in manufacturing 'true' tetrapods where each leg is the same size. However, new techniques are being developed that are increasing the number of tetrapods in a given sample to around 90 %, paving the way for more efficient cheaper solar power generation.

Generating more power for a given size of solar panel will clearly have a major impact on space missions. However, nano-technology is also attacking the power problem from a different direction, namely the ability to significantly increase power storage. In recent years, a new carbon-based material known as graphene has revolutionized our thinking about a wide range of electronic applications. In particular, the production of graphene-based ultracapacitors [68] for use in space-based power storage has a huge advantage over traditional batteries in that they have a higher power capability (more than ten times the capacity), live longer with less maintenance, operate over a wider range of temperatures, and are lighter and more flexible, making them very useful for compacting into a given spacecraft design.

Another technology being developed to generate the power necessary to operate a spacecraft is that of the Stirling radioisotope generator (SRG). This uses heat from radioactive decay to drive a Stirling engine which rapidly heats helium to drive a piston, the motion of which is used to generate electricity. SRGs produce four times as much power as radioisotope thermoelectric generators in operation at the present time (e.g. [19]). Typically, pellets of Pu-238 provide the radiation to heat the Stirling engine with a resulting output of around 140 W with about 30 % efficiency. SRGs can weigh around 30 kg but this is mitigated by the fact that the power is available continuously, without constraints from spacecraft orientation, distance from the Sun, or shadowing on a planetary surface. SRGs are in development for possible future use in missions to the outer

planets and their moons, where solar panels provide minimum power due to their distance from the Sun, and for planetary surface use, where maintaining operation at night (and avoiding freezing) is required.

These technologies show great promise but are many years short of being space-ready. However, once developed they will significantly enhance the capability and flexibility of a spacecraft to provide power which, in turn, will enhance the capability of any payload. More efficient power generation reduces the mass required for solar panels or increases the power generation for a given mass, while more capable energy storage allows for more flexible operations, e.g. longer ‘night-time’ operations when the spacecraft is out of direct sunlight, more on-board processing capability during operational downtime, and longer emergency life-time in the event of problems. Improving the power capability of a space mission reduces the impact of one of the major factors that constrains the design and sizing of a payload. For further discussion on spacecraft electric power systems see [Chap. 10](#).

6.5.2 Advanced Propulsion

Throughout the Space Age, we have relied almost entirely on chemical propulsion to put us into orbit or onto interplanetary trajectories, with the odd variant on liquid or solid propellant with occasional use of ion drives for propulsion or station-keeping. However, chemical propulsion is relatively inefficient, heavy, and of limited flexibility. Chemically propelled missions that are not in the ecliptic plane, go against the Earth’s orbital direction, or require complex trajectories, including returning to Earth, have to carry so much fuel that the launch costs climb sharply and the mass available to the spacecraft and payload is severely limited. Gravitational assists are frequently used to minimize the amount of fuel needed but these can add to the mission duration and lengthen the time to attain operational orbit. A number of alternative propulsion approaches have been developed and have even been demonstrated in flight, but they are a long way short of being the propulsion of choice. Each of them is unique and provides benefits and challenges for the payloads they carry.

Solar electric ion propulsion is a proven technology that can propel a spacecraft to velocities ten times as large as conventional chemical rockets with an equivalent amount of fuel. Specific impulse (see [Sect. 4.5.3](#)) is about 10 times that of chemical propulsion. The low thrust produced prohibits rapid accelerations but provides a relatively gentle platform for hosting a payload. The high efficiency of an ion propulsion system makes it a strong candidate for space missions with either a high-energy requirement or frequent or continuous maneuvering. An ion propulsion system was

flown with great success on Deep Space-1 (DS1), a technology demonstration mission of NASA’s New Millennium Program [38, 69] and on NASA’s Dawn mission to the asteroids Vesta and Ceres [70]. In the case of DS1 the ion thrusters were in continuous operation for almost 2 years and produced a velocity in excess of 4.3 km/s. Sensitive to the potential effects on a payload from the ion engine, DS1 also carried a set of diagnostic instruments designed to quantify the interactions of the ion propulsion system with the spacecraft. Diagnostic measurements that were made included the rate and extent of contamination around the spacecraft from the ion plume (Xe^+) and the sputtered material from the grid, the generation of electric and magnetic fields, and the density and energy of electrons and ions in the vicinity of the spacecraft. All of these interactions could potentially interfere with the working of the payload, contaminate measurements of the local environment or interfere with communications to ground stations. The diagnostic tests carried out over the course of the DS1 mission did not indicate any undue effects of the ion drive on the instrument function nor on their operation.

Another promising technology for high velocity space missions is the development of solar sails, also discussed in [Sects. 11.9.2](#) and [24.8](#). Solar sails are large, lightweight, reflective mirrors that use the light from the Sun as a form of propulsion [71]. The key property that defines the performance of the sail is its areal density (g/m^2). A low areal density can result in high accelerations, smaller sail size for a given mass constraint, or larger payload mass carried for a given sail size. High performance sails with areal densities $<1 \text{ g/m}^2$ can effectively defy gravity and enable a wide range of non-Keplerian, high-energy orbits and trajectories. Potential applications include sample returns, out-of-the-ecliptic orbits, polar ‘hover’ missions over planets or the Sun, and even interstellar missions. In 2011, the Japanese Space Agency (JAXA) flew a demonstration solar sail mission, IKAROS, with a square sail that was 20 m across the diagonal when fully deployed [36]. The sail was made from a 7.5-micron thick polyimide substrate and used embedded electrochromic panels to change the reflectance for attitude control. It was a relatively low performance sail, but it successfully tested both propulsion and attitude control for the first time in space for such a device. IKAROS also tested thin film solar cells, integrated into the solar sail film, with the ultimate goal of producing a dual solar sail, ion propulsion system. More capability can be obtained by increasing the sail size to achieve accelerations sufficient to enable high-energy missions. A major issue associated with using large solar sails for propulsion is the potential constraints on the payload operations. In order for a solar sail to be effective it must be as flat as possible to make maximum use of the incoming photons. This is accomplished by maintaining a tension on the film of the

sail and this is generally done by either attaching the sail to deployable booms or using the centrifugal force to deploy a circular sail with the spin of the spacecraft maintaining the sail tension. In the former scenario a 3-axis stabilized spacecraft is sufficient whereas a spinning, or dual spin, spacecraft is required in the latter. This feeds directly into the payload design. In addition, the space environment is pervaded by energetic charged particles, magnetic fields, plasma, dust, micrometeoroids, and debris, all of which can interact with the sail as it moves through the interplanetary medium. These interactions can have a number of potentially deleterious effects on mission operations and measurement contamination. A potential source of measurement perturbations is the charging of the sail material due to solar ultraviolet radiation [72] and plasma interactions. The accumulation of charge resulting from these processes can generate substantial potentials that then increase the interaction with the solar wind, changing the plasma environment around the spacecraft. Strategies for the measurement of the properties of the ‘undisturbed’ ambient medium need to be factored into the design of the payload: use of extendable booms, measurements made in the shade of the sail, etc.

The capability provided by these advanced propulsion techniques is exciting as new mission concepts can be devised which go far beyond those available with chemical propulsion (polar hovers over the Sun and the planets, interstellar missions, sample returns, rapid transit between Earth and Mars to name a few). The range of possible missions allows for innovative mission objectives with their associated impact on the payload. For further discussion on spacecraft propulsion systems see [Chap. 11](#).

6.5.3 Deployable Systems

Payload sizes are limited by the ability of the launch vehicle to get them into orbit. In addition to mass constraints, there are also length and volume constraints. The size of the payload is therefore limited and, as a result, so is the capability. For telescopes, space adds the advantage of zero-atmosphere to limit the ‘seeing’, while providing access to wavelengths not observable on the ground. However, the development of increasingly sophisticated adaptive optics schemes for ground-based telescopes means that the early advantages of observing from space with 0.5 and 1 meter-scale telescopes are no longer relevant, and certainly not an effective use of the money that it costs to launch them. In addition, as advancing technology enables observation of fainter or more distant objects or to higher resolutions, the need for larger and larger telescopes and antennas increases. The trick to overcome this apparent impasse is the application and implementation of deployable systems such as inflatable or extendable structures.

Examples of deployable structures include solar sails, large area optics, extendable instrument or support booms, solar concentrators, and large antennas. The development of all of these structures is ongoing to meet particular needs in the space community for military, commercial, and scientific purposes. Large area optics are of particular interest, as they will help overcome the current limitations of space-based telescopes and lead to significant improvements in the light gathering power with consequent improvements in resolution. Very large antennas have been proposed for a number of applications including larger bandwidth telecommunications, very long baseline interferometry, and space-based radar.

The obvious impact of such deployable systems on the payload design is the accommodation of the deployment mechanism, the additional support structure for the deployed system, and the control of, and communications between, the payload hub and the sensors, cameras or other instrumentation distributed throughout the deployed mass. A deployable structure on a spacecraft is usually a direct requirement designed to meet one of more of the mission objectives and as such drives the design of the payload. Decisions need to be made as to whether the deployment mechanism serves any useful purpose once the structure is deployed. In the case of some solar sail missions, for example, once the spacecraft has reached its desired operational location, the sail itself can be jettisoned. The torques created by a large deployed system, the space charge that it builds up, and the disturbance it makes in the medium it is traveling through, may all have a direct effect on the operation, measurements, or observations of the payload instrumentation. This presents a number of additional challenges in the design of the payload and its interface with the parent spacecraft.

6.5.4 Cryogenic Payloads

A range of new sensor technologies are being developed that enable the detection of individual photons resulting in significantly higher sensitivities and, as a consequence, the ability to provide high resolution simultaneously in space, time, and energy. These single-photon detectors will significantly advance the observational capability of telescopes in space. These sensors rely on taking standard calorimetric measurement techniques into the milli-Kelvin temperature regime. When photons impact the detector their energy is converted to heat that, with a significantly sensitive detector, can be measured. Typically, a measurable amount of heat requires a large number of incident photons that are in sparse supply for faint objects or brighter objects that need to be highly resolved in space or time. To detect single photons and thereby make efficient use of every photon, extremely sensitive calorimeters (micro-calorimeters) are required. Such

calorimeters must therefore be isolated from even the weakest sources of thermal background, from either space or the spacecraft itself. This is where cryogenics comes in. Since the 1980s, the Stirling cooler has been the most popular means by which to maintain a low temperature on board a space mission. Modern designs allow for temperatures of 50–80 K with two-stage Stirling coolers getting down to around 20 K. For a mission like the Planck observatory [73], which measured the microwave background at 2.7 K, a series of cooling systems were required to get to a background temperature of 0.1–0.3 K. This used a passive radiator to get to ~ 60 K, an H₂ Joule–Thomson cooler to get to 20 K, a Joule–Thompson mechanical cooler to get to 4 K, and then a dilution refrigerator to get to 0.1 K.

The goal is to get cryogenic temperature down to a few tens of milli-Kelvin. This can be achieved with adiabatic diamagnetic refrigerators (ADR), which have an advantage over standard dilution refrigerators, whose pipelines can get clogged and whose systems rely on the evacuation of the coolant (helium) into space, thereby requiring a significant mass of helium to be carried.

A class of detectors, known as ‘3D detectors’ [74] have been developed which operate in the tens of milli-Kelvin regime and allow single photon detections: hence the ability to simultaneously measure spatial, temporal, and spectral information. These cryogenic detectors rely on measuring the low-energy solid-state excitations resulting from single photon detections and so must be operated at temperature significantly below 1 K. Two examples of these micro-calorimeters are transition-edge sensors (TES) and superconducting tunnel junctions (STJ). These detectors have different operating principles but they are both photon counting detectors with energy resolutions, E/dE , of around 500, temporal resolutions of <1 ms, and spatial resolutions only limited by the size of the array that can be built. TES detectors are based on the sharp resistive transition of a thin superconducting film and are typically operated at temperatures below 0.1 K. The sensors are connected to a thermal bath that is maintained at a temperature a little lower than this operating temperature using an ADR. The absorbed photon produces a heat pulse that results in a transient decrease in the TES current, measured by a superconducting quantum interference device (SQUID). STJ detectors rely on the creation of quasi-particles in one or both of two thin superconductor layers separated by a thin insulating tunneling layer. STJs measure photon energies from the increase in tunneling current after absorption in one of the superconductor layers excites additional charge carriers above the superconducting energy gap. In both cases active cooling is required.

Cryogenic micro-calorimeter technologies are paving the way for major advances in the observation of astrophysical phenomena, particularly at X-ray energies. However, the accommodation of the requisite cryogenic system provides a

major challenge for the payload design. Typical ADR cryogenic systems have a mass of around 25–35 kg, comparable to a complete instrument. In addition, the reliance on magnetic fields for the cooling adds the need for shielding, which comes with an additional mass and size cost. Acknowledging this problem, ADRs are being developed which are small, with low mass and low power requirements and smaller less problematic magnetic fields, e.g. [75]. As micro-calorimeter instruments become more common and the potential science return increases, payloads will have to adapt to the increased reliance on cryogenic systems.

6.6 Conclusions

To design and size a payload, the mission objectives must be understood, along with how these flow down to the actual measurements to be made, instrument functional requirements, and instrument operations to be performed. This requires a detailed knowledge of how the payload will operate, how it interacts with the spacecraft, how it influences the mission operations, and how it is constrained by mission and spacecraft parameters (launch vehicle, fairing, orbit, structure, mass etc.). In addition, external factors such as cost and schedule may influence the payload design and ultimately lead to a change in the mission objectives. The crucial element in successful payload design is effective communication between the payload designers, spacecraft subsystem designers, and systems engineers. From early in the payload design to its final integration, testing, and launch, it is important to maintain constant communication between the payload development and the various factors feeding this development: science/operational objectives, mission capabilities, spacecraft design, technology implementation, calibration, cost, schedule, and risk assessment. From the early design concepts to the final payload, the flow of information between the various mission teams is crucial in keeping to cost and schedule, minimizing problems, and maximizing the final payload performance.

One of the most important aspects of payload design is the attention that must be paid to the various sets of requirements: *mission*, *spacecraft*, *performance*, and *measurement* (see Fig. 6.3). It is a careful definition of these requirements and their flowdown into the various subsystems that set the parameters for the final integrated payload. The requirements can be modified as the design develops, to take account of the many issues that crop up in the mission development, changing capability of the technologies, unforeseen design impasses, cost and schedule overruns, etc. However, these can be minimized by a robust and clear definition early in the design, maintaining cost and schedule targets and communicating any changes to the mission teams in a timely fashion.

This chapter has attempted to detail the wide array of considerations that factor into the design of a payload, whether it be a relatively simple transmitter in a geostationary orbit or a complex scientific suite of instruments heading for a distant planet. This is a complicated task and a single chapter overview of the payload design process is no substitute for the level of step-by-step detail required to build a successful payload and to integrate it into a successful mission. Several excellent texts exist that provide the necessary detail at the specific task level [2, 3, 65]. However, it is important to also understand the context in which payloads must operate and this has provided the focus for the present chapter.

Acknowledgments

Part of the research and work that supported this chapter was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. NASA Procedural Requirements 8705.4, *Risk Classification for NASA Payloads*, NODIS Library, 2004.
2. *Design Methodologies for Space Transportation Systems*, W. E. Hammond, AIAA Education Series, 2001.
3. *Space Missions Analysis and Design*, eds. W. J. Larson and J. R. Wertz, Space Technology Series, 1992, (Kluwer, Dordrecht)
4. *The Hubble Space Telescope mission, history, and systems*, L. L. Endelman, in International Congress on High-Speed Photography and Photonics, 19th, Cambridge, England, Sept. 16-21, 1990, Proceedings (A92-45101 19-35). Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1991, p. 422-441.
5. *NASA Solar Dynamics Observatory Mission website*: http://www.nasa.gov/mision_pages/sdo/spacecraft/
6. *The Interstellar Boundary Explorer (IBEX): Tracing the Interaction Between the Heliosphere and Surrounding Interstellar Material with Energetic Neutral Atoms*, P. C. Frisch and D. J. McComas, 2010, Space Science Reviews, DOI: [10.1007/s11214-010-9725-0](https://doi.org/10.1007/s11214-010-9725-0).
7. *The Microwave Anisotropy Probe (MAP) Mission*, C.L. Bennett, et al., 2003, ApJ, 583, 1.
8. *The Aeronomy of Ice in the Mesosphere (AIM) mission: Overview and early science results*, Russell, J. et al., 2009, J. Atmos. Solar-Terrestrial Phys., 71, 289.
9. *Voyager: The Grand Tour of Big Science*, Butrica, A. J., 1998, from Engineering Science to Big Science: The NACA and NASA Collier Trophy Research Project Winners, 251.
10. *The THEMIS Mission*, V. Angelopoulos, 2008, Space Science Reviews, 141, 5.
11. *The ULYSSES scientific payload*, Caseley, P. J. and Marsden, R. G., 1990, ESA Bulletin (ISSN 0376-4265), no. 63, 29.
12. *The Cluster Mission: ESA'S Spacefleet to the Magnetosphere*, Credland, J., Mecke, G., and Ellwood, J., 1997, Space Science Reviews, 79, 33.
13. *Apollo 11 Mission Report*, Mission Evaluation Team, NASA Manned Spacecraft Center, 1971, NASA SP-238, NASA, Washington DC.
14. *The Genesis Mission*, ed. Russell, C. T., 2003, Space Science Reviews, 105, Nos. 3-4. Kluwer Academic Publishers, Dordrecht.
15. *Stardust: Comet and interstellar dust sample return mission*, Brownlee, D. E. et al., 2003, Journal of Geophysical Research, 108, SRD 1-1, DOI [10.1029/2003JE002087](https://doi.org/10.1029/2003JE002087).
16. *Sample Return Mission to NEA: MUSES-C*, Fujiwara, A., Mukai, T., Kawaguchi, J., and Uesugi, K. T., 2000, Advances in Space Research, 25, 231.
17. *Lunar Reconnaissance Orbiter (LRO): Observations for Lunar Exploration and Science*, Vondrak, R., Keller, J., Chin, G., Garvin, J., Space Science Reviews, 150, 7.
18. *Spacecraft exploration of Mars*, Snyder, C. W., and V. I. Moroz, in Mars, Kieffer ed, U. of Arizona Press, 1992.
19. *The Cassini/Huygens Mission to Saturn*, Mitchell, R., 2000, Technical Report, Jet Propulsion Lab., California Inst. of Tech. Pasadena, CA 01/2000.
20. *The Galileo Mission*, Russell, C. T., 1992, Space Science Reviews, 60, 1.
21. *Scientific Results of the Viking Project*, Flin, E.A., 1977, Journal of Geophysical Research, 82, 3951.
22. *Global Mobile Satellite Systems: A Systems Overview*, eds. P. A. Swan and C. L. Devieux Jr., 2003, Kluwer Academic Publishers, Dordrecht.
23. *The Intelsat Global Satellite System*, Alper, J. and Pelton, J.N., Progress in Astronautics and Aeronautics Series, 93. AIAA.
24. Spaceway-1 Datasheet from Boeing.
25. Anik F2 Datasheet from Boeing.
26. *INMARSAT: Proceedings of the International Conference on Mobile Satellite Communications*, 1989, Blenheim Online Publications.
27. *The global positioning system*, Parkinson, S., 1996, American Institute of Aeronautics and Astronautics.
28. *A Beginner's Guide to GNSS in Europe*, EVP Europe, 1999, International Federation of Air Traffic Controller's Associations.
29. *The SPOT satellite remote sensing mission*, Chevrel, M., Courtois, M., Weill, G., 1981, Photogrammetric Engineering and Remote Sensing, 47, 1163.
30. *Eye in the Sky: Story of the Corona Spy Satellites*, eds. D.A. Day, J.M. Logsdon, and B. Latell, 1999, Smithsonian Books.
31. *Wings in Orbit Scientific and Engineering Legacies of the Space Shuttle, 1971-2010*, eds. W. Hale, H. Lane, G. Chapline, and K. Lulla, 2011, NASA Johnson Space Center.
32. Reference guide to the International Space Station. – Assembly complete ed., 2010, NASA document: NP-2010-09-682-HQ.
33. Soyuz: A Universal Spacecraft, Hall, R. and Shayler, D., 2003, Springer-Praxis, New York.
34. The Advanced Composition Explorer. E.C. Stone, A.M. Frandsen, R.A. Mewaldt, E.R. Christian, D. Margolies, J.F. Ormes, F. Snow, 1998, Space Science Rev., 86, 1.
35. NASA GOES History website: <http://goes.gsfc.nasa.gov/text/history/goes/goes.html>.
36. *First Solar Power Sail Demonstration by IKAROS*, Mori, O., et al., 2009, 27th International Symposium on Space Technology and Science, July 5-10 2009, Tsukuba, Japan.
37. *Principal VASIMR Results and Present Objectives*, Glover, T. W. et al., 2005, SPACE TECHNOLOGY AND APPLICATIONS INT.FORUM-STAIIF 2005: 22nd Symp Space Nucl. Powr Propuls. AIP Conference Proceedings, 746, 976.
38. *Deep Space One: NASA's first Deep-Space technology validation mission*, Rayman, M. D., and D. H. Lehman, 1997, Acta Astronautica, 41, 289.

39. *Jupiter Icy Moons Orbiter (JIMO): An Element of the Prometheus Program*, 2004, Technical Report, JPL Publication 04-16; 982-R06933
40. New Millennium Program Space Technology 7 website: <http://nmp.jpl.nasa.gov/st7/>
41. New Millennium Program Space Technology 5 website: <http://nmp.jpl.nasa.gov/st5/>
42. *The CubeSat: The Picosatellite Standard for Research and Education*, R. Nugent, R. Munakata, A. Chin, R. Coelho, Dr. Jordi Puig-Suari, 2008, AIAA Space 2008 Conference, San Diego California.
43. *In-flight validation of the formation flying technologies using the ASPIICS/PROBA-3 giant coronagraph*, Vivès, S.; Lamy, P.; Levacher, P.; Venet, M.; Boit, J. L., 2008, Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter. Edited by Oschmann, Jacobus M., Jr.; de Graauw, Mattheus W. M.; MacEwen, Howard A. Proceedings of the SPIE, 7010, 70103R.
44. *Autonomous Formation Flying for the PRISMA Mission*, Gill E., D'Amico S., Montenbruck O., 2007, AIAA Journal of Spacecraft and Rockets, 44, 671.
45. *Commercial Space Transportation Study*, 1997, Commercial Space Transportation Study Alliance, United States Aerospace Corporation, <http://www.hq.nasa.gov/webaccess/CommSpaceTrans/>.
46. *The SOHO Mission*, Fleck B., Domingo, V. and A.I. Poland, Solar Phys., 162, 1.
47. Project Apollo: The Tough Decisions, Seamans, Robert C. Jr., 2005, Monograph in Aerospace History, No. 37, NASA SP-2005-4537.
48. *LCROSS Science Payload Ground Development, Test and Calibration Results*, Ennico, K.; Colaprete, A.; Wooden, D.; Heldmann, J. L.; Kojima, G.; Shirley, M., 2008, 39th Lunar and Planetary Science Conference, (Lunar and Planetary Science XXXIX), LPI Contribution No. 1391, p.1474.
49. *Mars Global Surveyor Mission: Overview and Status*, Albee, A. L., Palluconi, F. D.; Arvidson, R. E., 1998, Science, 279, 1671.
50. *The Sun*, D. Alexander, 2010, Greenwood Guides to the Universe, Greenwood.
51. *Temporal variations of strength and location of the South Atlantic Anomaly as measured by RXTE*, Fürst, F., Wilms, J., Rothschild, R. E., Pottschmidt, K., Smith, D. M., Lingenfelter, R., 2009, Earth and Planetary Science Letters, 281, 125.
52. *Storms from the Sun: The emerging science of space weather*, Carlowicz, M. J. and Lopez, R. E., 2002, The Joseph Henry Press, Washington, DC (USA).
53. *A Critical Overview on Spacecraft Charging Mitigation Methods*, Lai, 2003, IEEE Transactions on Plasma Science, 31, 1118
54. *Atomic Oxygen Effects on Spacecraft Materials*, Banks et al. 2003, NASA/TM—2003-212484
55. *Cleanrooms and associated controlled Environments*, ISO-14644-1
56. *Space Product Assurance Cleanliness and Contamination Control*, European Space Agency, ECSS-Q-70-01A, 2002.
57. *Product Cleanliness levels and Contamination Control Program*, MIL-STD-1246C, 1994.
58. *The detection of organic contamination of surfaces by infrared spectroscopy*, European Space Agency, ECSS-Q-70-05A, 2005.
59. *Space shuttle glow observations*, Banks, P. M., Williamson, P. R. & Raitt, W. J., 1983, Geophys. Res. Lett. 10,118.
60. *Spacecraft Thermal Control Handbook Volume II: Cryogenics*, Martin Donabedian, 2003, Aerospace Press Series, Aerospace Press.
61. *The Wide-Field Infrared Survey Explorer (WISE)*, Duval, V.G., Irace, W.R., Mainzer, A.K. and Wright, E.L., 2004, Optical, Infrared, and Millimeter Space Telescopes. Edited by Mather, John C. Proceedings of the SPIE, 5487, 101.
62. *The Spitzer Space Telescope Mission*, Werner, M.W. et al., 2004, ApJ Supp., 154, 1.
63. *Information theory, Inference and Learning Algorithms*, MacKay, D. J. C., 2003, Cambridge University Press.
64. *State of The Art Lossless Image Compression Algorithms*, Sahni, S., Vemuri, B.C., Chen, F., Kapoor, C., Leonard, C., and Fitzsimmons, J., 1998, IEEE Proceedings of the International Conference on Image Processing, Chicago, Illinois, 948.
65. *Cost-Effective Space Mission Operations*, Boden, D. G. and Larson, W. J., 1996, College Custom Series, McGraw-Hill, New York.
66. *Calibration and flight of the NRL EIT CalRoc*, Newmark J. S. et al., 2000, Proc. SPIE Vol. 4139, p. 328-339, Instrumentation for UV/EUV Astronomy and Solar Missions, Silvano Fineschi; Clarence M. Korendyke; Oswald H. Siegmund; Bruce E. Woodgate; Eds.
67. *Performance of CdSe tetrapods-gold as nanostructure electrochemical materials in photovoltaic cells*, Liu, T.-Y., Eukel, J.A., Bagaria, H., Wong, M.S., Pasquali, M., and Schmidt, H.K., 2009, in Photovoltaic Specialists Conference (PVSC), 34th IEEE, 2074.
68. *Graphene-Based Ultracapacitors*, Stoller, M.D., Park, S., Zhu, Y., An, J. and Ruoff, R.S., 2008, Nano Lett., 8, 3498.
69. *Results from the Deep Space 1 Technology Validation Mission*, Rayman, M.D., Varghese, P., Lehman, D.H. and Livesay, L., 2000, Acta Astronautica 47, 475.
70. *Dawn: A mission in development for exploration of main belt asteroids Vesta and Ceres*, Rayman, M.D., Fraschetti, T. C., Raymond, C. A. and Russell, C. T., 2006, Acta Astronautica 58, 605.
71. *Solar Sailing: Technology, Dynamics and Mission Applications*, McInnes, C.R., 1999, Springer-Praxis, Chichester.
72. *Microscopic Approach to Analyze Solar-Sail Space-Environment Effects*, Kezerashvili, R.Y. and Matloff, G.L., 2009, Advances in Space Research, 44, 859.
73. *The FIRST/Planck Mission. Cryogenic systems - Current Status*, Collaudin B. and Passvogel, T., 1998, Proc. SPIE Vol. 3356, p. 1114-1126, Space Telescopes and Instruments V, Pierre Y. Bely; James B. Breckinridge; Eds.
74. *Quantum Calorimetry*, Stahle, C.K., McCammon, D. and Irwin, K.D., 1999, Physics Today, 52, 32.
75. *A miniature continuous adiabatic demagnetization refrigerator with compact shielded superconducting magnets*, Duval, J.-M., Cain, B.M. and Timbie, P.T., 2004, Millimeter and Sub-millimeter Detectors for Astronomy II, Edited by Jonas Zmuidzinas, Wayne S. Holland and Stafford Withington Proceedings of the SPIE, 5498, 802.

Vincent L. Pisacane

A space system consists of a complex set of synergistically related components that together satisfy a coherent set of requirements derived from a set of needs. The objective of systems engineering is to design, develop, deploy, operate, and dispose of a system that meets the user's needs, defined in terms of technical or performance specifications and constraints such as cost, schedule, and risk that constitute a set of system-level requirements. Requirements at all levels should be interpreted to include both technical requirements and constraints. Consequently, systems engineering is the interdisciplinary systematic and concurrent development and verification of a product or service to satisfy the system-level requirements. Often, a system is viewed as an independent entity, but in actuality, it will interact with other systems and exist within the context of larger or super systems. Systems are generally categorized hierarchically as consisting successively of segments, elements, subsystems, assemblies, subassemblies, and parts. Segments, elements, and subsystems have explicit hierarchical requirements derived from system-level requirements. Consequently, each is often called a system as well. In this chapter the use of the term system is often intended to apply equally to segment, element, or subsystem.

Systems engineering was devised at the Bell Telephone Laboratories in the 1940s and was further developed by the United States Department of Defense, NASA, and other entities into a more formal discipline [1, 2]. Technological advances that today permit the distributed development of increasingly complex systems that require progressively increasingly specialized skills have necessarily increased the importance of and dependence on systems engineering to achieve a successful outcome. NASA and ESA continue to lead the way in developing space systems engineering

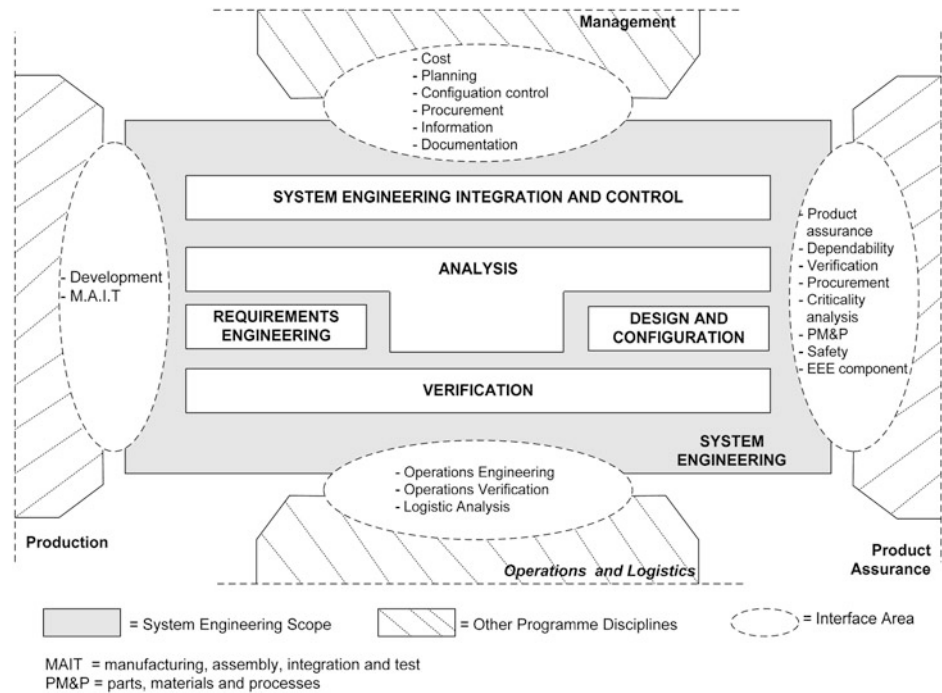
practices through documented practices, training, and standards, while the broader field of systems engineering is being developed by organizations such as the Electronics Industry Association (EIA), International Council on Systems Engineering (INCOSE), International Standards Organization (ISO), and International Electrotechnical Commission (IEC).

Systems engineering can be described as a formalized and disciplined approach to the development, deployment, utilization, and disposal of a system that satisfies specific needs, formalized by a set of needs and technical requirements or specifications within the bounds of stringent constraints. Thus, needs includes the characteristics of the potential work product and concomitant constraints. The successful realization of systems engineering is a system that satisfies the stated needs and balances the technical requirements and constraints, with the latter often being cost, schedule, and risk. The role of the systems engineer begins with an understanding of the needs, development of the system requirements, and ends when the system is disposed of or decommissioned. In modestly sized projects there may be one systems engineer, but in large, geographically distributed projects a team may be necessary. The lead systems engineer must be cognizant of the overall engineering activities and play a critical role in developing the systems engineering management plan (SEMP). His/her responsibilities include leading the development of the system architecture, defining interfaces, allocating requirements among the subsystems, evaluating trade-offs, assessing risks, assuring verification and validation, leading system-level reviews, assessing subsystem reviews and tests, heading the configuration control board, and assuring proper documentation. Inadequate planning and failure to adhere to a formal systems engineering management plan can lead to a failure to meet the requirements and specifically to cost and schedule overruns. Figure 7.1 illustrates

V. L. Pisacane (✉)

United States Naval Academy, Annapolis, Maryland, USA

Fig. 7.1 Systems engineering functions and boundaries, from [3]



the relationships between systems engineering and the other requisite functions in the development of a system.

Figure 7.2 illustrates the systems engineering engine that drives the system development process. Steps 1 through 9 represent the sequence of tasks required to execute a project, while steps 10 through 17 are crosscutting tools for carrying out the processes to complete the development.

Using, for demonstration purposes, selected NASA projects developed in the 1970s and 1980s, Fig. 7.3 illustrates the benefits of expenditures on the definition phases to reduce cost overruns. The definition costs are the actual costs expended in steps 1 through 4 in Fig. 7.2 that include Phase A (Concept and Technology Development) and Phase B (Preliminary Design and Technology Completion) of the development life cycle. The target costs are the estimated costs and the actual costs are the realized costs subsequent to the definition costs (i.e. subsequent to Phase B). The data in Fig. 7.3 suggest that the optimum expenditures during the definition phases are in the vicinity of 15 % of the total project estimate; however it is noteworthy that few data points sit beyond 10 %. This appears to support the contention in systems engineering of the benefit from having well-defined phases with clearly defined and vetted needs and hierarchical requirements.

Considerable information is available to help define and assist in the application of systems engineering. ISO/IEC-15288:2008 [6] defines a set of processes and terminologies applicable to any level in the hierarchy of the development process. For software development, the standard ISO/IEC-12207:2008 [7] may also be useful. Handbooks that describe the role and practices of systems engineering are

available from NASA/SP-2007-6105 [4], INCOSE [8], and ECSS-E-HB-10 [9].

A systems engineer must provide direction from technical, management, and leadership perspectives. It is generally accepted that a systems engineer will have a foundational background in one of the traditional engineering disciplines (e.g., mechanical, electrical, industrial, computer engineering), experience as a lead engineer in the development of hardware subsystems, and additional training in systems engineering. A review of practicing systems engineers would show a remarkable diversity in their education, experience, and career paths. Nonetheless, displayed in Table 7.1 are the characteristics that a systems engineer should possess, recognizing that no one single individual could enjoy all of those identified.

7.1 Concepts in Systems Engineering

In developing a sophisticated and complex system it is important that all participants work together in a coherent and synergistic manner. This is achieved through the Systems Engineering Management Plan (SEMP), which documents how the technical and engineering activities are to be carried out in a fully integrated manner. The SEMP generally consists of ten sections, as illustrated in Table 7.2. Its objective is to define the approaches, procedures, resources, organizational structures, levels of responsibilities, and commensurate levels of authority used to address all aspects of each of the life cycles of the project. The systems engineer typically has the responsibility to develop the

Fig. 7.2 The systems engineering engine, from [4]

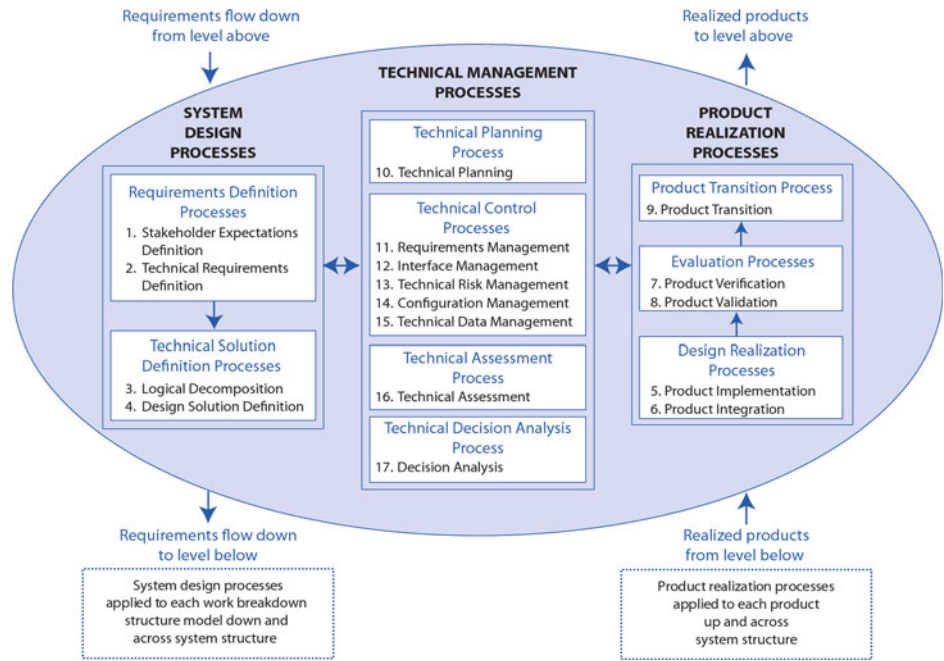
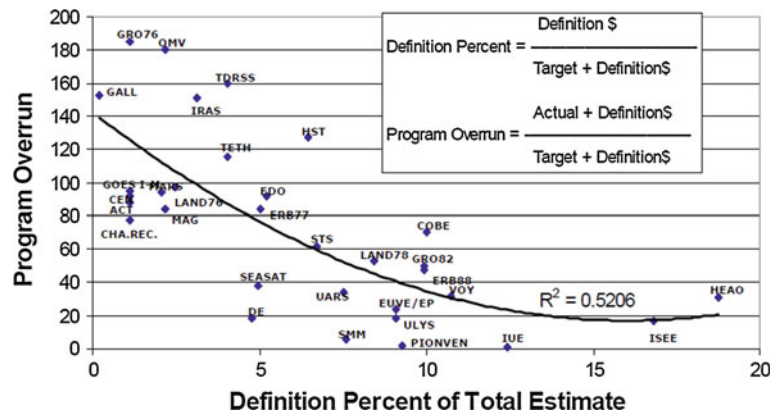


Fig. 7.3 Correlation of definition phase expenditures and project cost overruns from 32 NASA Programs, from [5]



SEMP but the document should integrate inputs from all participating organizations and groups, especially project management. Specifics of the SEMP are given in NASA NPR 7123.1A, [10] and ECSS-E-ST-10C [3].

Several procedural models are used to develop hardware- and software-intensive space systems, including the most popular waterfall and spiral development models. The waterfall model comprises a phased progression of functions leading to the realization of the system. A number of variants exist, with typical phases of the model illustrated in Fig. 7.4. The waterfall model is most appropriate for projects with a limited number of realizations for which the requirements are well defined at the beginning of the project and remain independent over time. Advantages include the emphasis on requirements and design prior to implementation that are intended to reduce rework, and that the phases have defined entry and exit criteria so that progress

can be compartmented and the status quantitatively assessed. Deficiencies of the waterfall model are that requirements may not be sufficiently defined, requirements may change over the development period, and errors made in the requirements and design phases may not be discovered until late in the project when remediation is costly. The application of the waterfall model to software development is discussed in Sect. 16.4.1.

The spiral development model consists of a series of phases that repeat themselves with increasing definition and capabilities until the project has satisfied all the requirements and the system development is complete. A four-phase spiral development model is shown in Fig. 7.5, in which design, implementation, integration, and testing, are iterated with increasing resolution in each cycle. Also illustrated is that spiral development may be viewed as the waterfall model applied recursively. It is suited to projects

Table 7.1 Desirable characteristics of systems engineers

Capabilities	Technical	Management	Leadership
General		Able to solve problems Accepts responsibility Can-do attitude Disciplined Good bidirectional communicator and listener High energy Learns independently Makes informed decisions Takes calculated risks Thinks critically Understands systems engineering Understands the role of the systems engineer Can delegate Listens effectively Self confident Open minded	
Specific	Diverse technical skills Engineering or physics education Experience with simulations Hands on hardware experience Knows electrical engineering Strong technical experience in one or more relevant areas Willingness to learn new technologies	Acts as a member of the supporting teams rather than above them Appreciates importance of programmatic performance, schedule, and costs Delegates commensurate responsibilities and authority Promotes teaming Understands key elements of program management Uses meetings efficiently and effectively	Has support of upper management Focuses on the big picture Willing to delegate both responsibility and authority Appreciates importance of team cohesion Exhibits personal accountability and expects it of others Promotes teaming Sets and displays high standards

in which the initial requirements are not well established or may change as a function of time, or when it is appropriate to deploy a system with limited capability on an interim basis. Otherwise, the spiral model is expected to result in a longer schedule and be more costly than the waterfall model. It is most appropriate for software development, where limited capability implementations may be more easily deployed and recursively enhanced, as opposed to hardware-intensive systems. The application of the spiral model to software development is discussed in Sect. 16.4.1.

The Interface Control Document (ICD) is a tool to assure the development of a system according to a strict set of requirements. The ICD is a structured means of communicating information about interfaces among design teams. It is a formal document that is often the key element in a contract and the basis for legal action when it is believed that the product does not satisfy the specifications. The ICD should be signed by at least one representative of each affected subsystem and the systems engineer. Specifications in the ICD are derived from the system-level requirements that are often specified in an Interface Requirements Documents (IRD). This document defines the interfaces between systems

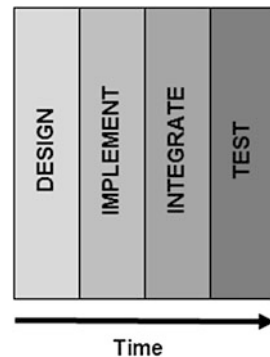
and not the characteristics of the relevant systems. Typically, the ICD describes the interfaces to the lowest level; e.g., voltages and their tolerances, connector types, and data formats. A well-defined ICD allows the system under development to utilize only a simulator of the specified inputs and the measurement of its outputs to carry out interface and performance tests. The ICD is developed during the design phases and is formalized subsequent to the critical design review. Once signed, changes should be adjudicated by the configuration control board. Employment of the ICD is described in NASA/SP-2007-6105 [4] and ECSS-M-ST-40C [11]. A typical ICD format is given in Table 7.3.

Trade or trade-off analyses are used in systems engineering to assure that the optimum system or subsystem is developed. They are a systemic approach to balancing requirements, especially performance, cost, schedule, and risk. A well-recognized axiom in systems engineering is known as the *systems engineer's dilemma* that states

- To reduce cost/schedule at constant risk, performance must be reduced.
- To reduce risk at constant cost/schedule, performance must be reduced.

Table 7.2 Contents of a typical systems engineering management plan

Section	Content
Purpose and scope	Defines project and provides purpose, scope, and overview of the SEMP
Applicable documents	Provides references and project documentation
Technical summary	Provides executive summary of the mission and system to be developed. Includes the system description, system structure with work breakdown structure, integration procedures, system and development constraints, and references to non-technical plans
Technical effort integration process	Identifies engineering procedures used to produce the deliverables with sufficient detail to guide the development teams satisfying cost, schedule, and performance objectives. Describes organizational structures, roles, responsibilities, authorities, and technical management tools to support technical integration
Common technical processes implementation	Identifies common technical processes and requirements to meet exit criteria for each project phase and project objective. Includes processes for requirements definition, technical solution definition, technical planning, technical control, technical assessment, technical decision analysis, product transition, evaluation and design realization. Inherent in these are management of requirements, interfaces, technical risk, configuration, and technical data and system verification and validation
Technology insertion	Describes methods for identifying and assessing key technologies and the risks and criteria for inserting them
Additional SE functions and activities	Identifies processes for functions not previously identified such as safety, reliability, human factors, logistics, maintainability, quality, operability, and supportability
Integration with the project plan resource allocation	Identifies how technical requirements will be integrated with the project plan to determine allocation of resources, including cost, schedule, and personnel, and how changes to the allocations will be coordinated
Waivers	Identifies any approved waivers from organizational requirements
Appendices	Includes glossary, acronyms and abbreviations, and information given separately for convenience

Fig. 7.4 Waterfall development model

- To reduce cost/schedule at constant performance, higher risks must be accepted.
- To reduce risk at constant performance, higher costs and a longer schedule must be accepted.

To effectively carry out a trade-off analysis requires a thorough understanding of the requirements of the project. Based on this, various alternative mission concepts, architectures, and subsystems can be proposed, each of which can be modeled throughout its life cycle phases. This is generally accomplished recursively. Sometimes it is appropriate to reach a decision by identifying and comparing the advantages and disadvantages of each alternative. However, a more formal approach is often used based on a priori weights that sum to unity for a set of a priori defined metrics. The metric is

a measure for quantitatively assessing a desirable or undesirable characteristic. Positive weights can be used for desirable metrics and negative values for undesirable metrics. Once the metric quantities are determined from modeling each candidate alternative, they may be combined to determine an overall assessment. The weighted metrics are generally combined linearly or multiplicatively, although other techniques can also be used. The weighted linear approach is the most commonly used method. It is advantageous when a low value for one or more metrics for mission critical functions cannot be offset by other high value metrics. During the trade-off analysis it is often advisable to perform a sensitivity analysis to examine whether changes to the requirements would make an important difference in the outcomes. If so, these trade-offs can be offered as potential alternatives for consideration by program management.

Technology readiness levels (TRL), discussed in [Chap. 2](#) and illustrated in [Fig. 2.10](#), are the quantitative assessments of the maturity of the technology of a system, subsystem, or assembly. TRLs are a useful tool to facilitate the systematic process of transitioning a technology from the research and development phase to the operational phase. As such, they identify technical risk by enumerating the number of levels that a technology must survive to be considered mature and operationally vetted. The more numerous the lower TRLs planned for a system, indicating reliance on less mature

Fig. 7.5 Spiral development model

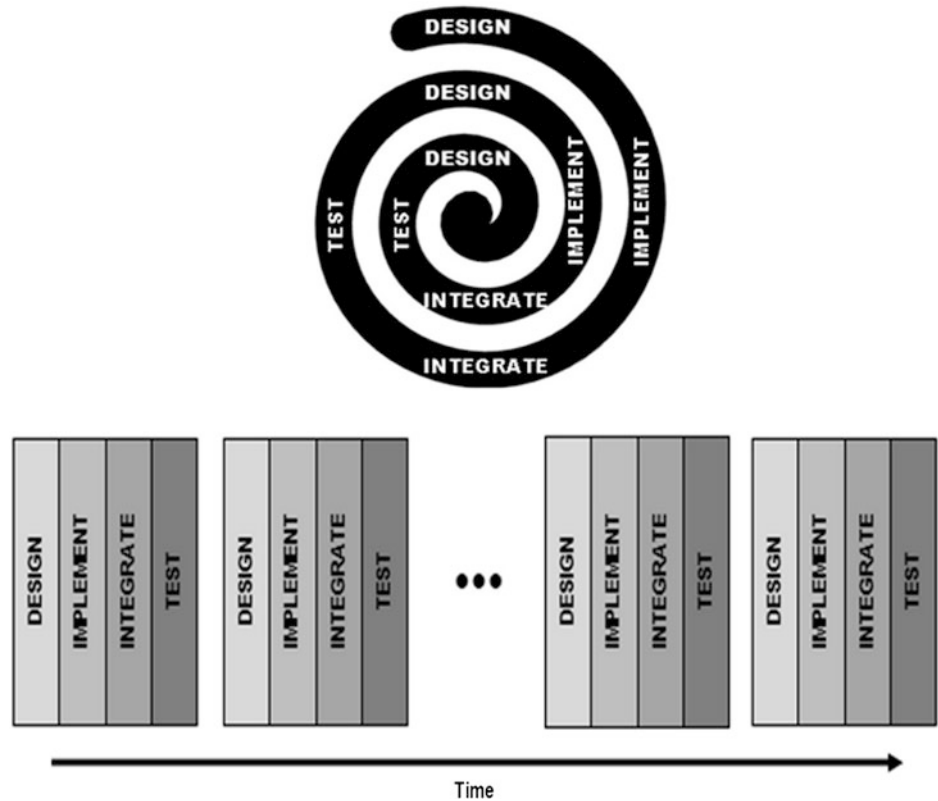


Table 7.3 Sample template for an interface control document

1. Introduction
1.1 Purpose
1.2 Scope
1.3 Applicable Documents
1.4 System Overview
1.5 Operational Agreement
2. Interface Requirements
2.1 Physical/Mechanical
2.2 Radio Frequency Interfaces
2.3 Commands and Data Interfaces
2.4 Physical and Mechanical Interfaces
2.5 Electrical Interfaces etc.
3. Interface Verification
3.1 Reliability
3.2 Quality Assurance
3.3 Tests
Appendices
A Figures
B Tables
C Definitions, Acronyms, and Abbreviations
D References

technologies, the more likely is it that the project may suffer schedule delays or cost over-runs.

A critical resource for a space system is a property such as component mass, end-of-life power, data storage, average and peak data rates, schedule, and cost. Throughout the development of a system, a critical resource has several assessments that include an initial estimate, a current estimate, and an allowed magnitude. The magnitudes of the critical resources often tend toward the allowed magnitude as the project matures. It is important to track critical resources so that they can be evaluated to determine if corrective action is anticipated or needed for those that are approaching their allowed magnitudes. The margin or contingency of a critical resource is the difference between the allowed magnitude and the current estimate, with the percent margin or contingency being the margin divided by the current estimate times 100. The growth is the difference between the current estimate and the initial estimate, with the growth factor being the current estimate divided by the initial estimate. Mass growth factors and power growth factors of 1.1 to 1.2 are not unusual. An anticipated contingency factor is an a priori estimate of the typical growth of a critical resource whose value is based on its current state of maturity. Anticipated contingency factors for typical growths for mass as a function of the resource's maturity are given in Table 7.4. The less mature the item the greater is the contingency factor.

Table 7.4 Anticipated contingency factors

Maturity	Anticipated contingency factors
Off the shelf or measured	1.05
Minor modification of an existing design	1.07
Modification of existing design	1.10
New design, calculated mass	1.15
New design, with thoughtful estimated mass	1.20
New design, with high uncertainty in estimated mass	1.30

7.2 Project Development Life Cycle

As introduced in Chap. 2, major space systems are generally developed according to a project life cycle model that consists of distinct phases, each of which has identified and categorized activities that must be completed to produce specified deliverables. Partitioning the project life cycle into phases has the advantage that each phase is built on a completed prior stage with specified deliverables that have been independently reviewed and evaluated. In addition, decision points between phases represent go or no-go decision points for management. The project life cycle phases as defined by ECSS and NASA are listed in Table 2.2.

Pre-Phase A or Phase 0 generally consists of formulating a program and identifying potential missions, supported by advanced studies comprising analyses, simulations, and some limited research or exploratory developments. The primary objectives are to formulate relevant missions, mission goals, system-level requirements, preliminary concept of operations, preliminary costs, and preliminary schedules and identify technologies that need to be developed. Often an interdisciplinary science working group with broad participation is used to help establish the programmatic scientific objectives.

Phase A comprises the conceptual formulation of a specific project and may be in response to a solicitation, to meet a user's needs, or with the intent to develop an unsolicited proposal. The primary activities include defining the overall mission; refining the needs; developing system-level requirements; developing one or more conceptual designs; identifying needed research, exploratory developments, and long-lead items; defining an initial concept of operations; and developing an initial schedule and cost estimate.

Phase B consists of selecting the optimum design from those identified in Phase A, if there are several, and carrying out preliminary designs of the mission and system and its deployment, operation, and disposal. These preliminary designs are characterized by sufficient detail on which to base firm estimates of performance, operational characteristics, risk, cost, and schedule. The major go or no-go decision for a project is generally made after the completion of Phase B.

Phases C and D are generally integrated. In Phase C the design is finalized as a critical design in sufficient detail so that fabrication can follow, substantiated by critical design reviews at the subsystem and system levels. Following a successful system critical design review, the design is frozen and fabrication is initiated in Phase D, which comprises fabrication, assembly, integration, verification, and validation at the subsystem and system level followed by deployment of the system in its operational environment.

Phase E constitutes operation of the system and involves its maintenance and the continued training of the operators and maintenance of the ground segment.

Phase F includes decommissioning and documentation of the system's overall performance.

Figure 7.6 expands on Fig. 2.9 and illustrates the NASA life cycle phases and the phased major events and life cycle gates that need to be successfully met. The reviews identified are discussed in Sect. 1.4.

The purpose, typical activities, products, and reviews at the end of each phase are given in more detail in Table 7.5.

Different life cycle strategies are generally employed for non-flight programs such as basic and applied research and advanced technology development, depending on the nature, complexity, and the technology readiness level. These projects often follow the model of formulate, approve, implement, and evaluate, as illustrated in Fig. 7.7.

7.3 Needs and Requirements

The first stage in determining a specific set of requirements or specification is a needs assessment or needs analysis. This identifies the characteristics of the work product and associated constraints, and it must be independent of any proposed solution, focusing instead on identifying strengths, weaknesses, opportunities, advantages, constraints, and threats. The assessment should be documented, and accompanied by any analyses performed with a set of coherent performance needs and constraints, along with measures of effectiveness or performance. The needs assessment should be formally vetted by the stakeholders.

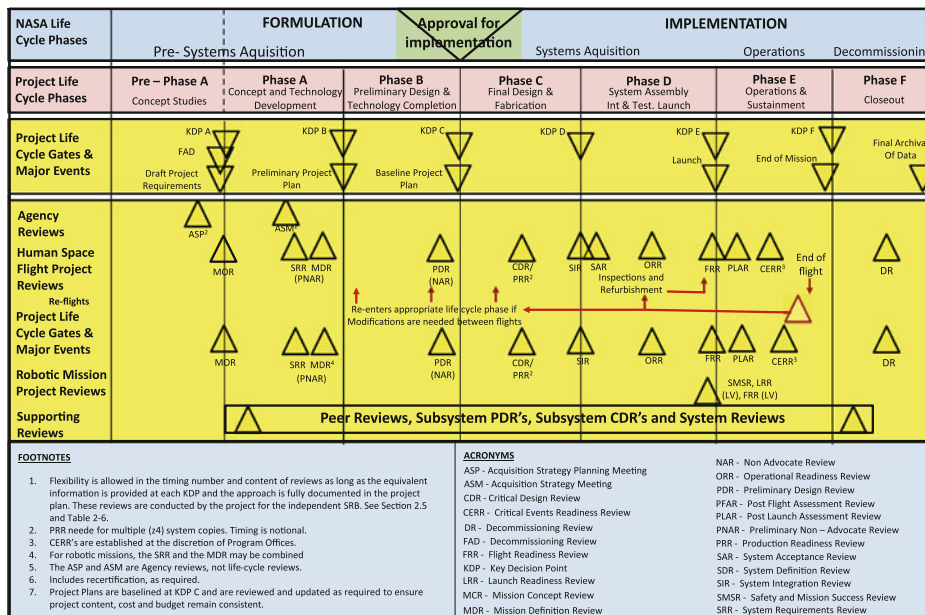


Fig. 7.6 NASA life cycles with major events and life cycle gates, from [10]

The needs assessment must specify the *what* but not the *how*. Once the needs have been established they can be transformed into system-level requirements that include technical requirements or specifications and constraints.

The system-level requirements initially developed from the needs assessment specify as precisely as possible a description of the desired properties of the system and their constraints independently of how they are to be realized. They, as all requirements, should be definable, realizable, measurable, and verifiable. Inadequate, incomplete, imprecise, unclear, conflicting, and unrealizable requirements will invariably lead to difficulty in the later phases of the project. Each requirement should be quantitative and accompanied by tolerances, justifications, assumptions, and be observable and, therefore, measurable.

The system-level requirements should be categorized as technical requirements, such as performance, operation, reliability, safety, environmental, human factors, maintainability, usability, testability, supportability, producibility, etc. and constraints, such as cost, schedule, regulations, policies, and so forth.

An appropriate set of requirements establishes the basis of agreement among participants, reduces rework activities, provides a basis for estimating cost and schedule, and provides a baseline for verification and validation. Characteristics of good requirements include

- **Clarity**—are they concise, unambiguous, and consisting of one requirement per statement?
- **Completeness**—does each one stand alone, are all assumptions given with tolerances, and is the sum sufficient to satisfy the needs?

- **Compliance**—are the requirements at the correct level (system, segment, element, subsystem, etc.) and independent of the implementation?
- **Consistency**—are the requirements non-contradictory and is the terminology consistent?
- **Traceability**—are the requirements at a lower level traceable to those at a higher level, for example, can the subsystem requirements be related to specific system-level requirements?
- **Correctness**—are the technical requirements technically feasible, with specified assumptions, and are relevant analyses provided?
- **Interfaces**—are all interfaces clearly defined?
- **Reliability**—are reliability requirements clearly defined?
- **Verifiability**—are the requirements sufficiently specific to be verifiable?

Development of detailed system-level requirements is initiated by a functional decomposition and analysis activity. Functional decomposition and analysis is the systematic recursive process that identifies, characterizes, and inter-relates the hierarchical functions that a system must carry out in order to satisfy the system-level requirements. Each system-level requirement is analyzed to identify the specific functions that need to be carried out. Explicitly excluded is how the functions are to be performed. These characteristics are generally achieved by recursive and iterative procedures that involve a thorough understanding of the needs, system-level requirements, and the technical state-of-the-art. This process will also often reveal requirements that would not have been otherwise identified.

Table 7.5 NASA life cycle phases, modified from [4]

Phase	Purpose	Products	
Formulation	Pre-Phase A Concept Studies	Produce a broad spectrum of ideas and alternatives for missions from which new programs/projects can be selected Determine feasibility of desired system Develop mission concepts Draft system-level requirements Identify potential technology needs	Feasible system concepts in the form of simulations, analysis, study reports, models, and mockups
	Phase A Concept and Technology Development	Determine feasibility and desirability of a suggested new major system Establish an initial baseline Develop mission concept Develop system-level requirements Identify needed technology developments Pass conceptual design review	Mission objectives System-level requirements Mission conceptual design Mission architecture Mission operational plan Conceptual subsystem requirements Technology requirements Schedule and cost
	Phase B Preliminary Design and Technology Completion	Establish an initial baseline Develop preliminary mission architecture Develop preliminary operational plan Develop preliminary system requirements Develop preliminary subsystem requirements Develop a preliminary subsystem design Develop schedule and cost Demonstrate feasibility Pass preliminary design review	Mission objectives System-level requirements Mission preliminary design Mission architecture Mission operational plan Preliminary subsystem requirements Status technology development Schedule and cost
Implementation	Phase C Final Design and Fabrication	Develop final mission architecture Develop final operational plan Develop final system requirements Complete detailed design of system Complete detailed design of subsystems Demonstrate maturity of needed technologies Pass critical design review Fabricate, integrate, and test subsystems.	Mission objectives System-level requirements Mission critical design Mission architecture Mission operational plan Critical subsystem requirements Status of technology developments Schedule and cost Tested subsystems
	Phase D System Assembly, Integration and Test, Launch	Integrate and test the system Verify it meets requirements Validate it meets needs Launch	Verified and validated system ready to be deployed Deployed system
	Phase E Operations and Sustainment	Initiate the mission Implement mission operations plan Sustain operations support	Operational system
	Phase F Closeout	Decommission system Dispose of system	Disposed system

The functional architecture, generally developed first through the top-down process, specifies and describes each system-level function to be carried out. This is followed by the creation of functional flow block diagrams that illustrate the sequence and timelines for all system functions. Each block diagram should represent a single function to be performed, described by its function, inputs, outputs, constraints, and interfaces. A consistent numbering scheme should be used to relate hierarchically each functional block diagram. A timeline analysis is used to relate the sequencing and duration of each of the functions. Once accomplished at the top level, the functional decomposition and analysis for each identified function can be developed at sequentially

more detailed levels, with an appropriate hierarchical numbering scheme. The results of this activity are more detailed system and lower-level requirements or specifications, the systems architecture with component requirements for each function, and a concept of operations, or ConOps, that collectively describes the system architecturally and operationally. Figure 7.8 illustrates an example of the functional flow block diagram to the third level.

The interfaces between functional blocks can be represented by an n-square diagram; see Fig. 7.9. Functional blocks from the functional flow block diagrams are placed along the diagonal representing the functions to be carried out. Off-diagonal blocks represent the functional interfaces

Fig. 7.7 NASA technology development life cycles, from [12]

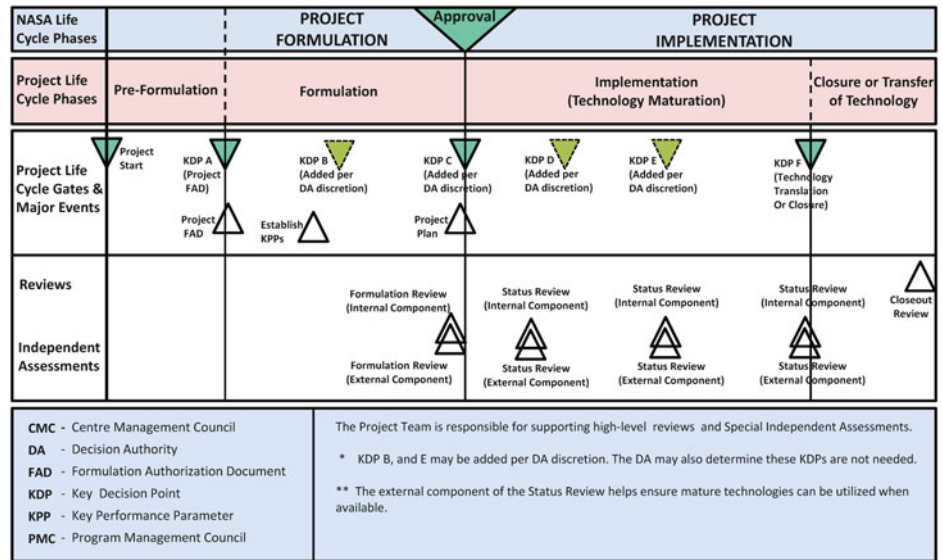
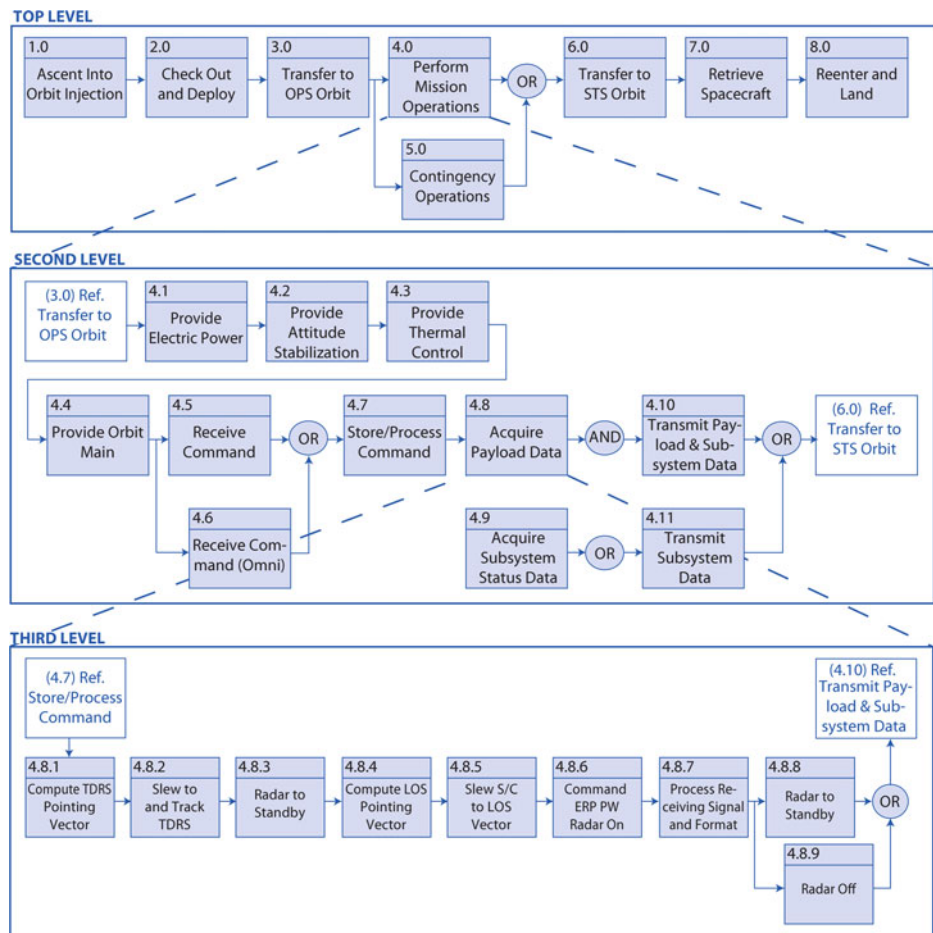


Fig. 7.8 Example of a functional flow block diagram, from [4]



between the functional blocks, where the flow of information between functions is shown by the direction of the arrows. For any block, e.g., F_i , the column represents the inputs from all other blocks and the row represents the

outputs to all other blocks. If there is no interface between two functional blocks, the block that represents their interface is blank. The n-square diagram assists in assuring that all interfaces are appropriately assigned.

F_1	$F_1 \rightleftarrows F_2$	$F_1 \rightleftarrows F_3$...	$F_1 \rightleftarrows F_n$
$F_2 \rightleftarrows F_1$	F_2	$F_2 \rightleftarrows F_3$...	$F_2 \rightleftarrows F_n$
$F_3 \rightleftarrows F_1$	$F_3 \rightleftarrows F_2$	F_3	...	$F_3 \rightleftarrows F_n$
...
$F_n \rightleftarrows F_1$	$F_n \rightleftarrows F_2$	$F_n \rightleftarrows F_3$...	F_n

Fig. 7.9 N-square diagram where $F_i = \text{Function } i = 1 \text{ to } n$

Once the functional decompositions and functional analyses are completed it is possible to begin the design solution process. This involves transforming the system and subsystem requirements into a realized system through the remaining life cycle procedures identified in Sect. 7.2.

Figure 7.10 illustrates the flowdown of requirements to the subsystem level. An allocated requirement (or simply requirement) is one that is directly related to a higher level requirement. Thus, an allocated requirement can be directly traced back to a specific higher level requirement. In contrast, a derived requirement is a requirement that is not explicitly related to a higher level requirement but is imposed independently to satisfy a higher level requirement. Although not explicitly indicated by the higher-level requirement, it is inferred from the context of the higher level requirement. Derived requirements are generated during the formulation phases and must be established to fully specify the system. Consequently, derived requirements are typically not part of the initial requirements documents. However, once the derived requirements are established it is necessary to allocate them to the appropriate subsystems in order to fully specify the design requirements. An example of an allocated requirement is the mass allocated for a subsystem imposed to satisfy a top-level requirement on mass. Alternatively, voltages available from the power bus would not be expected to be traceable to a higher-level requirement but would be imposed as a derived requirement for other subsystems.

It is important that the characteristic of each work product is traceable to each subsystem requirement, which is in turn traceable to specific system-level requirements, and that each system-level requirement is traceable to the user needs. Bidirectional traceability of requirements will ensure that each of the lowest level requirements is required to satisfy the user's needs. In addition, if it becomes necessary to change a requirement, each aspect affected can be appropriately identified. These relationships are often identified in a bidirectional requirements traceability matrix (BRTM). The BRTM provides a convenient way to assuredly locate and identify the impact of any proposed change to all affected entities. Backward traceability can verify that the work product is consistent with the requirements and needs. The BRTM is also useful in developing and carrying out test plans and assisting in verification and validation

assessments. BRTMs can be developed using the procedure described in Fig. 7.11, utilizing a variety of tools including requirements management software, databases, spreadsheets, tables, or hyperlinks. In its simplest form, it consists of a spreadsheet with vertical columns of hierarchical numbering system, requirements, and sources of the requirement.

Essential characteristics for requirements traceability to be consistent and complete are

- The requirements traceability matrix should identify all the system-level requirements that include the technical requirement and constraints.
- The requirements traceability matrix should list all the lower-level requirements that include the technical requirement and constraints.
- The requirements identified should be identical to those identified in other documents specifying requirements, specifications, and constraints.
- Each system-level requirement should be linked to at least one lower-level requirement.
- Each requirement at a lower-level element should be linked to a system-level requirement.
- If a requirement at a lower level is not linked to a requirement at a higher level, this requirement should be separately justified.

7.4 Technical Assessment

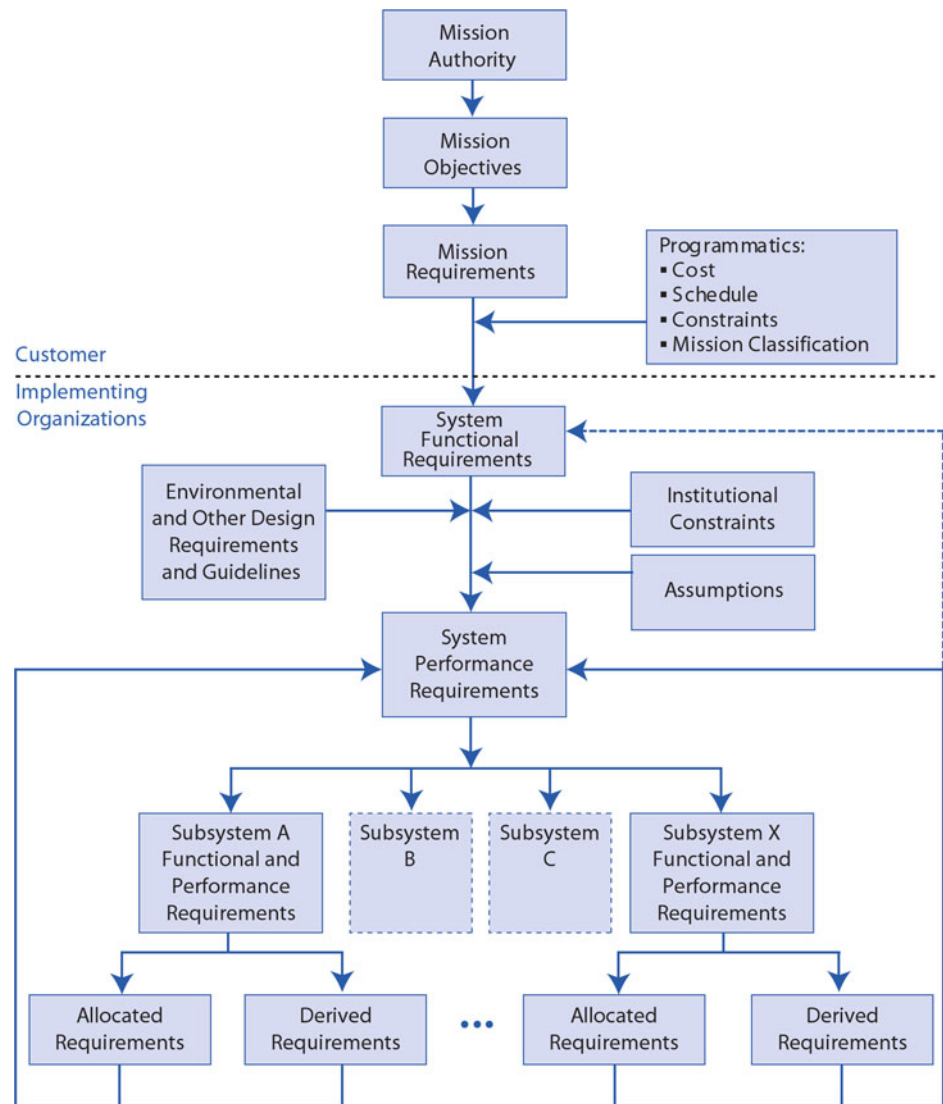
Monitoring throughout the project life cycle is essential to assure that the technical requirements are being achieved. Technical assessment is generally done independent of assessing fiscal status. Technical assessment is generally accomplished by technical reviews at the subsystem, element, segment, and system levels. These reviews provide effective mechanisms by which to communicate among participants and demonstrate that particular objectives or milestones have or have not been achieved. In particular, a review

- Provides an independent technical assessment of status
- Assures that interfaces are understood
- Promotes communication among participants
- Formalizes and documents that milestones have been achieved
- Identifies outstanding issues
- Provides affirmation to management of the status of the project.

To assure a competent and complete review it is important that

- A review panel is chartered to provide the assessment
- The members of the panel have no conflict of interest and are knowledgeable of the technologies to be covered

Fig. 7.10 Requirements flowdown, from [4]



- Specific purposes and objectives of the review are clearly defined
- Review is scheduled at a time when presenters are ready and have adequate time to prepare
- A dry run of presentations is held
- Agenda should allow sufficient time for each presentation
- Material provided to panel and interested parties is clear, succinct, and consistent and sufficiently prior to the review
- Chairperson meets with project management and reviewers to confirm the purposes, objectives, and the protocol
- Chairperson controls meeting
- Forms are available for members and audience to identify issues and questions
- Tentative action items are identified and documented during the meeting
- Splinter meetings on specific topics may be directed by chairperson
- Tentative action items, questions, and comments are integrated by the review panel into a manageable number of formal action items
- At the conclusion the action items are identified, discussed, possibly amended and assigned to a particular individual with a due date for completion
- The review panel submits a definitive report.

Development of a system is not complete without rigorous testing and assessments to assure that the implementation is consistent with the requirements and intended use. A system-level test and evaluation plan should be initiated in Phase A, developed concurrently with the system-level requirements, and assessments should be specifically linked to the requirements. The plan should then be iterated with increasing specificity throughout the subsequent phases of

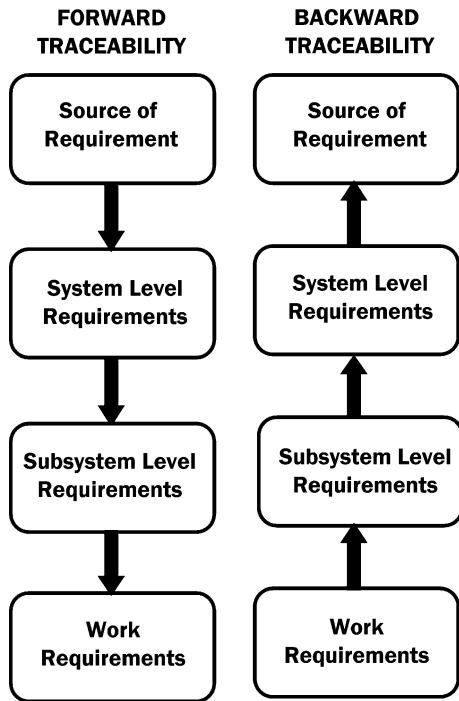


Fig. 7.11 Bidirectional requirements traceability process

the project. As the system-level requirements are hierarchically decomposed into segment, element, subsystem, assembly, and subassembly requirements, appropriate tests need to be concurrently developed for each level to demonstrate that those requirements are satisfied. Thus, these tests assure compliance of the development with regard to requirements specified for that level. It is critical that each test plan be reviewed to assure that the test article will be at an appropriate stage to be tested, the test will demonstrate its required functionality and assess the requirements, the test facility is appropriate with sufficient safeguards, the personnel to carry out the test are qualified, the appropriate data are to be collected, and the test results will be fully documented. These test results should be reported on in the subsequent technical reviews. The types of reviews encountered in the development of a space system follow in Fig. 7.12 and are time sequenced.

The reviews in Fig. 7.12 are

- CDR Critical Design Review: A multi-disciplinary technical review to assess whether the system is ready to proceed into system fabrication, integration, and test
- CERR Critical Events Readiness Review: Assesses the project's readiness to execute a critical activity
- ConR Continuation Review: Assesses the status of a project to recommend for or against continuation

- DR Decommissioning Review: Assesses the readiness of the system to be safely decommissioned and disposed
- FRR Flight Readiness Review: Assesses the readiness of the system and support activities to achieve a successful launch and follow-on flight operations
- MCR Mission Concept Review: Assesses whether the mission concept will satisfy the overall requirements and needs
- MDR Mission Definition Review: Assesses whether the mission concept and architecture are technically feasible and match available resources
- ORR Operational Readiness Review: Assesses whether the system characteristics, operational procedures and documentation, and capabilities of the operators and support systems can support a successful operation
- PDR Preliminary Design Review: A multi-disciplinary technical review that assesses whether or not the system design is sufficiently complete to conclude Phase B and proceed into Phase C
- PFAR Post-Flight Assessment Review: Assesses launch anomalies and issues and recommends actions to improve future launches
- PLAR Post-Launch Assessment Review: Post-launch assessment of the readiness of the spacecraft systems to proceed with normal operations
- PRR Production Readiness Review: Assesses the readiness to efficiently begin production by review of the production plans
- PTR Periodic Technical Review: Assesses technical progress and status
- SAR System Acceptance Review: Assesses the completeness of the deliverables by reviewing the design, documentation, and test results to assure the requirements are satisfied
- SDR System Definition Review: Assesses the system architecture and requirements flowdown
- SIR System Integration Review: Assesses whether the system, element, or subsystem is ready to be integrated
- SRR System Requirements Review: Assesses the adequacy of the system functional and performance requirements
- TRR Test Readiness Review: Assesses the readiness of the test facility, capabilities of the support personnel, and appropriateness of the test procedures to determine if a test can be safely carried out.

Verification and validation, known as V&V; independent verification and validation known as IV&V; and verification, validation, and accreditation known as VV&A; are

Fig. 7.12 Systems engineering technical reviews, from [10]

Flight Systems and Ground Support Projects				Basic & Applied Research	Advanced Technology Development	Institutional Projects*
	Human	Robotic	AO - Driven			
FORMULATION	Concept Studies (Pre - Phase A)	▽ MCR □ FAD	▽ MCR □ FAD	▼ Down Select # 1	Prep Portfolio Process	□ SEMP
	Concept & Technology Development (Phase A)	□ SEMP □ PP ▽ SRR	□ SEMP □ PP ▽ SRR	□ SEMP □ PP	Approval Process	▽ PTR ☞ ConR
	Preliminary Design & Technology Completion (Phase B)	▽ SDR	▽ MDR	▼ Step 2 Select	Evaluate Proposals	□ SEMP ☞ ConR
					Solicit. Rev. / Recom. Proposal for sel.	▽ PTR ☞ ConR
						▽ PTR
						▽ PTR
IMPLEMENTATION	Final Design & Fabrication (Phase C)	▽ CDR/PRR ▽ SIR ▽ TRR	▽ CDR/PRR ▽ SIR ▽ TRR	▽ CDR/PRR ▽ SIR ▽ TRR	Initiate Fund for Invest.	Technology Readiness Level Maturation
	System Assembly Integration, Test & Launch (Phase D)	▽ SAR ▽ ORR ▽ FRR ▽ PLAR	▽ SAR ▽ ORR ▽ FRR ▽ PLAR	▽ SAR ▽ ORR ▽ FRR ▽ PLAR	Monitor Perform	Key Performance Parameter Enhancements
	Operations & Sustainment (Phase E)	▽ CERR ▽ PFAR	▽ CERR ▽ PFAR	▽ CERR	Update Portfolio	▽ PTR ☞ ConR
	Closeout (Phase F)	▽ DR	▽ DR	▽ DR	Comm. Results	▽ PTR ☞ ConR
					Monitor Perform Metrics	
						Execute Project Plan

END OF MISSION

LEGEND	
	Management Life- Cycle Phases
	Product- Line Life- Cycle Phases
▼	Major Management Review/Control Gate
▽	Engineering / Technical Review
☞	Continuous Review
□	Documents
FAD	Formulation Approval Document
PP	Project Plan
SEMP	System Engineering Management Plan
CDR	Critical Design Review
CERR	Critical Events Readiness Review
ConR	Continuation Reviews
DR	Decommissioning Review
FRR	Flight Readiness Review
MCR	Mission Concept Review
MDR	Mission Design Review
ORR	Operational Readiness Review
PA	Portfolio Approval
PDR	Preliminary Design Review
PFAR	Post- Flight Assessment Review
PLAR	Post- Launch Assessment Review
PRR	Production Readiness Review
PTR	Periodic Technical Review
SAR	System Acceptance Review
SDR	System Definition Review
SIR	System Integration Review
SRR	System Requirements Review
TRR	Test Readiness Review

important processes to establish that the requirements have been met and that the system should meet the user's requirements and needs. IV&V differs from V&V in that the IV&V evaluation is carried out by personnel independent of the system development.

Verification is the process that evaluates whether or not the products of a life cycle phase satisfy the conditions imposed at the start of that phase. Verification can be carried out throughout the development life cycle and is

intended to identify deficiencies, redundancies, and discrepancies relative to the requirements. Activities undertaken include inspections of components and products, peer reviews, modeling, and testing. Testing is an activity by which a system is activated under specified conditions and the results are observed and documented so that an evaluation can be made. It is essential that the capability to adequately test a system be designed into it during the design phase.

Validation is the process that evaluates whether or not the system will serve its intended purpose. Consequently, it can only be carried out at the completion of the development when the completed system is available to be evaluated. Strategies for validation include system-level testing, alpha testing, and beta testing. Alpha testing is sometimes employed after acceptance testing in order to further validate the system. When utilized, it is generally carried out by the developers to identify any residual technical issues and further validate the system's operational functionality. Beta testing may be employed for systems when it is deemed important to further identify any technical or operational issues that may be encountered by the user after the systems has been deployed in the operational environment. Thus, this testing is generally performed by accommodating users. Compliance testing with respect to the project requirements was discussed earlier. In addition, compliance testing with respect to industry standards may be undertaken if the system is to interface with other systems. Standards compliance tests help to assure the interoperability of the system.

V&V are particularly important for assuring the integrity of software-intensive systems, especially those with extensive embedded software. V&V plans are developed early as part of the systems engineering process and define the goals, objectives, techniques, and documentation of the assessments. Planning and execution of the programs are generally carried out by V&V engineers who are not involved in developing the system. Thus, they are sufficiently independent to appropriately plan and implement the V&V programs.

Accreditation is the confirmation by a convening authority that the system will work as defined. This may be based on the documentation of the V&V programs, but may also be based on an independent assessment chartered by the convening authority.

7.5 Software Systems Engineering

As space systems become more complex, their subsystems have correspondingly increased in complexity with more reliance on complex firmware and software components. Sometimes the physical and logical boundaries between software and hardware can be uncertain. In many cases, the software is the most complex and challenging aspect of a subsystem or system. To efficiently and effectively develop firmware and software it is important to adopt a software systems engineering process. This is described in detail in [Chap. 16](#) and is discussed here briefly because software is an ever-increasingly critical aspect of satisfying the requirements within schedule and budget. Perceived as 'easy to change', software may provide flexibility to

overcome hardware deficiencies, especially late in the development cycle.

Software systems engineering and systems engineering have evolved essentially independently, but to successfully develop a system with important hardware and software components it is necessary to integrate these processes. Not adopting a systems engineering approach to software development may result in software that does not satisfy its requirements, does not integrate well with its hardware host or other elements of the system, is not easily changeable, and may be difficult to test, update, and document. Suffice it to say that software systems engineering follows a scenario similar to that described here for the system of which it is a component.

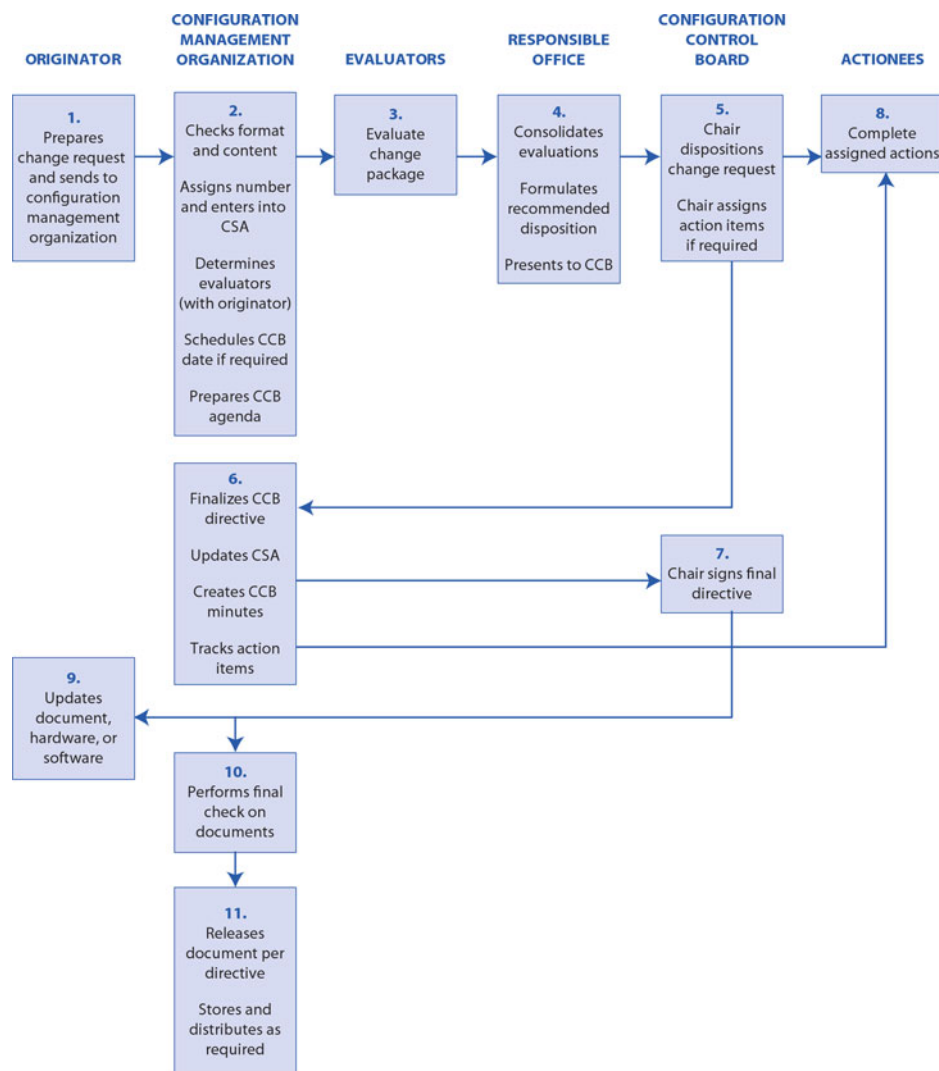
The international standard ISO/IEC-12207:2008 [7] describes software life cycle processes that can be used to acquire or develop software products or the software components of a system. The software engineering standard ECSS-E-ST-40C [13] and the software product assurance standard ECSS-Q-ST-80C [14] are both based on ISO/IEC-12207 [7]. NASA's software engineering requirements are described in NPR 7150.2A [15]. For a detailed description of software system engineering processes the reader is referred to [Chap. 16](#).

7.6 Configuration Management

Configuration management is the discipline that assures that a system conforms to its requirements and that its documentation sufficiently and accurately describes its functional and physical characteristics. The level of detail of the documentation should be sufficient to troubleshoot, repair, update, and replicate the system. A configuration management plan should be initiated at the beginning of Phase C. Configuration management control is generally instituted after the design has been frozen, subsequent to the critical design review in Phase C. Configuration management consists of five elements

- *Configuration planning and management*—Initiated by the development of a strategic plan that specifies the total configuration management activity, fully communicates it to all participants, and assures that participants comply with all provisions.
- *Configuration identification*—The structured process to identify the items to be under configuration management, the required documentation, and the change control authority.
- *Configuration change management*—The procedure to be followed to propose, adjudicate, resolve, and approve changes to the configuration. [Figure 7.13](#) illustrates a typical procedure for effecting orderly changes.

Fig. 7.13 Typical configuration change control procedure. Image: NASA [4]



- *Configuration status accounting*—Provides the status of the configuration of each item, such as historical configuration documentation and status of changes, waivers, discrepancies, and proposed changes. It provides the single authoritative source for baseline definitions.
- *Configuration verification*—Verifies that all changes have been incorporated, that the required documentation is complete and accurate, and that all requirements have been satisfied.

Finally, redlining is the procedure of making changes to documentations, drawings and documents by clearly marking up a controlled set of documentation instead of issuing new documentation. New documentation, especially detailed drawings, requires time to be completed, verified for accuracy, and approved. Redlining works well to reduce schedule delays in effecting changes but should not be used to circumvent verification and approval procedures.

7.7 Systems Engineering Organization

Technology management is described in [Chap. 21](#) and overall project management is described in [Chap. 22](#). Here the system engineering organization structure is succinctly discussed. The organization to assure effective systems engineering should be defined in the systems engineering management plan. It should be tailored to the specific project undertaken and communicated to all participants. For the systems engineering function to be effective and efficient, the organizational structure should have the following characteristics

- Clearly defines roles, responsibilities, and controls
- Delegates responsibility and commensurate control
- Establishes streamlined reporting procedures
- Facilitates bidirectional communications
- Standardizes and simplifies reporting and documentation

- Minimizes extraneous and repetitive reporting and documentation
- Promotes teaming.

For additional information on project organization and management the reader is referred to ECSS-M-20B [16] and NASA 7120.5E [17].

7.8 Risk Management

Effective risk management is critical to achieving mission success. Risk is defined as the potential inability to achieve a particular requirement and the potential for unexpected adverse outcomes. Risk management is the adoption of systematic procedures to identify and reduce risks to an acceptable level. Early detection is important in order to reduce the programmatic impact of a potential risk. Risks to a system include cost, schedule, and technical. Cost and schedule risks relate to the inability to satisfy the budget and meet the schedule. Technical risks relate to not satisfying the technical requirements. More specifically, risks may include such factors as criticality, national or international importance, reflight opportunities, financial investment, loss of life, and other relevant factors. A particular mission may include instruments or elements at different risk levels. However, no item should be dependent on another item that is at a higher risk level.

A risk management program is a systematic phased approach to identify, analyze, plan, track, and control potential risks while documenting and communicating them to the relevant parties, as illustrated in Fig. 7.14. At any phase in the sequence, it may be necessary to return to an earlier phase if new information leads to the identification of a new risk or changes a characteristic of a previously identified risk.

Risk identification may be based on expert opinion, lessons learned from implementing similar systems, test results, engineering analyses, or hazard analyses. The output from the identification phase includes the identification and description of each risk, its potential causes, priority, and consequences. Hazard analyses utilizes one of three approaches to identify potential failures

- *Failure Mode Effects Analysis (FMEA)*—An inductive or bottom-up approach that assumes a failure or defect at the lowest level and assesses its effects. It is used most advantageously early in the design process and should be reported on in design reviews to identify potential design weaknesses.
- *Failure Mode Effects Criticality Analysis (FMECA)*—Adds the probability of occurrence and the potential severity of failure to the FMEA process.
- *Fault Tree Analyses (FTA)*—A deductive top-down approach that identifies a potential failure and then identifies all events that could give rise to the failure.

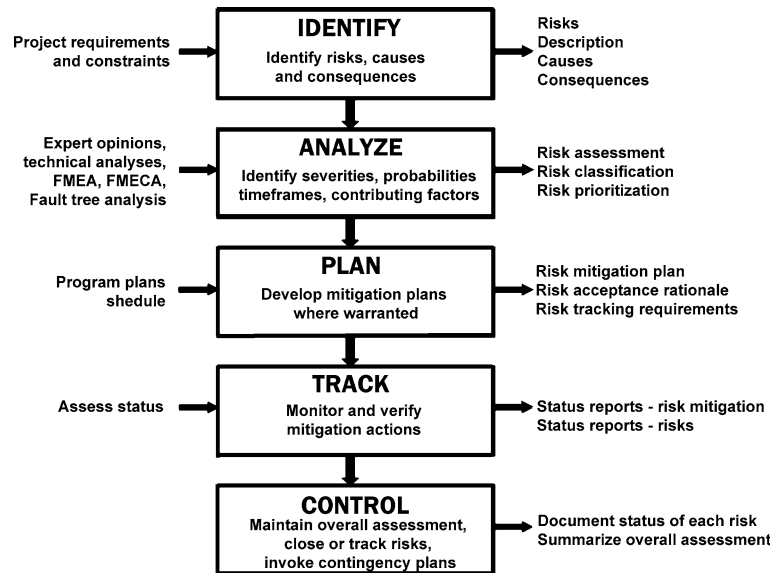
The analysis phase consists of evaluating the risks and their consequences, severities, probabilities of occurrence, timeframe of occurrence, contributing factors, and relative criticality. Several approaches can be used to prioritize risks. The most common is to decompose the risk into two components: the probability of occurrence and the potential consequences if it should occur, as illustrated in Fig. 7.15. The measures for each may be quantized as low, medium, or high, or as continuous, as from 0 to 100 or 0 to 10. Low impact and low probability of occurrence risks in the lower left hand corner have the lowest priority and are often accepted. High impact and high probability of occurrence risks in the upper right corner are of critical importance, have the highest priorities, and receive the most attention. Low impact and high probability of occurrence risks in the upper left corner and high impact low probability of occurrence risks in the lower right corner are of moderate to high importance and should be addressed as appropriate, depending on the impact and cost to mitigate. A specific instantiation of this approach often used in NASA projects is illustrated in Table 7.6 where the numbers from 1 to 7 indicate a relative prioritization with number 1 having the highest priority. Outputs of the analysis phase include the detailed assessment, classification of risks by type, prioritization by severity, and potential mitigating actions.

Risk planning relates to the development of a formal risk management plan to address each risk based on the appropriate analyses. Each plan includes provisional decisions on risk control, identification of observables and their thresholds, methodologies for documenting observables, proposed actions when observables exceed their thresholds, and assignment of responsibility for tracking risks. Observables are measurable parameters that provide information on the severity and probability of the risk. Possible decisions on risk control include: eliminate, mitigate, research, watch, or accept. The outputs of the planning phase are risk management plans for each risk that describe the procedures to be followed for mitigation.

Risk tracking is the procedure that monitors the status of the risks and action taken. The observables for each risk are tracked and reported on as identified in the risk management plan. If an observable exceeds a threshold; e.g., if an insignificant risk evolves into a significant risk or a new risk is identified, management is notified and appropriate action taken. Outputs of the tracking phase are reports on the status of the severity of the risk, probability of occurrence, and the mitigations undertaken.

Risk control is exercised to ensure that the risk management plan is appropriate, is being carried out, deviations are assessed and adjudicated, and up-to-date assessment of the overall risk and the status of each risk are provided to management. The outputs of the control phase are the

Fig. 7.14 Technical risk management procedures



identification of each risk by category, status of the risk mitigation activities, and the overall assessment of risk for the mission.

Having an a priori classification of acceptable risks provides a structured methodology that can be used as guidance to carry out risk management. Table 7.7 illustrates four classifications of risks used by NASA for payloads, subsystems, elements, or systems. Recognizing a priori the relative risk that is acceptable can help provide a more uniform approach to the risk mitigation.

The institution of fault protection that includes building into the system the capability to detect faults and institute responses can reduce the risk to space missions. While this function can be established cooperatively in the spacecraft and ground segments, it is usually instituted autonomously in each segment separately that places the segment in one of several planned safe modes. When this occurs, human intervention from either the spacecraft or ground segment (or both) can assess the situation and take corrective action. However, for critical events or the need for operational continuity, it may be necessary for the each segment to autonomously take corrective action by transitioning to back-up subsystems or an alternate mode of operation.

7.9 Cost and Schedule

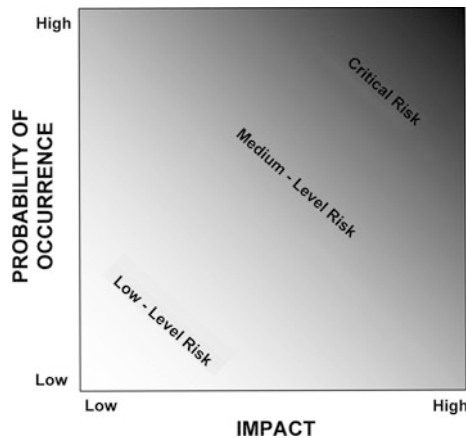
Cost and schedule are intrinsically linked to the engineering development by direct association with the activities that need to be carried out. An important tool in defining the work to be done is the work breakdown structure (WBS). This is the decomposition of the work to be done into distinct activities, starting at the system level and decomposing the activities hierarchically to the segment, element,

subsystem, assembly, and subassembly levels as appropriate. Each activity should have identifiable criteria to start the activity and measurable outputs at completion. Defined for each activity are: dependent and precursor activities; deliverables with final and intermediate milestones, together with their completion dates; required resources in personnel, purchases, subcontracts, and facilities; and unique issues such as specialized training needs and significant contracts. As a result of defining the activities in this manner, a schedule and budget can be determined at each level. This is generally accomplished with considerable feedback among levels in order to assure consistency. The activities at each level can be represented into a schedule by Bar or Gantt charts, milestone charts, or network diagrams. In defining the schedule at each level, attention must be directed to the critical paths and floats for each activity. Free float is the time that an activity can be delayed without affecting other activities. Path float is the time along a path in the schedule that activities can collectively be delayed without affecting the final delivery date. The critical path (or paths) are paths that have zero or near-zero path float. With specification of costs associated with each activity at each level, budgets allocations can be determined at each level of the work breakdown structure.

Design to cost is an important concept that is of increasing interest as budgets are receiving increased scrutiny where non-fixed cost contracts have historically been the norm. Previously, performance, schedule, and cost were prioritized in that order, with cost often an afterthought to the technical staff. The design engineer was often independent of the budgeting process and the monitoring of cost during the execution of the project. This is no longer the case as the design engineer is now typically equally accountable for performance, schedule, and cost. To make

Table 7.6 Risk matrix

Consequence	Likelihood estimate				
	Improbable	Unlikely to occur	May occur	Probably will occur	Likely to occur
Catastrophic	4	3	2	1	1
Critical	5	4	3	2	1
Moderate	6	5	4	3	2
Negligible	7	6	5	4	3

**Fig. 7.15** Risk categorization

this effective it is important that costs be allocated at the lowest reasonable WBS level and that expenditures be tracked at the same WBS level.

Earned value management is a methodology to more accurately assess the status of the schedule and cost of a project. When an assessment is to be made three costs are determined: budgeted cost of work scheduled, budgeted cost of work performed, and actual cost of work performed. The budgeted cost of work scheduled (BCWS) or planned value is the sum of the budgeted cost of the activities that should have been completed at the time of the assessment. The budgeted cost of work performed (BCWP) is the budgeted cost of the activities that are actually completed. Two rules are often applied for activities that have been started but are incomplete. The first approach for the BCWS and BCWP is to use a percentage of the budget based on the percentage of the work scheduled and performed. The second approach is to use 50 % of the budgeted costs for each until the activity is completed. The actual cost of work performed (ACWP) is the actual cost incurred and is generally determined from the allocation of costs to specific activities, as on time cards or cost accumulation system. This necessitates having a system in place to faithfully capture costs by activities to the required level. The total budget is the budget at completion (BAC) and the schedule

length is the schedule at completion (SAC). From these quantities the following estimates can be made at any time in the schedule

- The cost variance (CV) is given by the difference between BCWP and the ACWP, with a positive sign indicating that the project is under budget and a negative sign the project is over budget.

$$CV = BCWP - ACWP \quad (7.1)$$

- The schedule variance (SV) is given by the difference between BCWP and BCWS, with a positive sign indicating that the project is ahead of schedule and a negative sign the project is behind schedule.

$$SV = BCWP - BCWS \quad (7.2)$$

- The percent complete (PC) is the fraction of the work completed and is given by the BCWP divided by the BAC.

$$PC = \frac{BCWP}{BAC} \quad (7.3)$$

- The cost performance index (CPI) represents the cost efficiency and is given by the BCWP divided by the ACWP where favorable is >1 and unfavorable is <1.

$$CPI = \frac{BCWP}{ACWP} \quad (7.4)$$

- The schedule performance index (SPI) is a measure of how far the project is ahead (SPI > 1) or behind schedule (SPI < 1) and is given by the BCWP divided by the BCWS.

$$SPI = \frac{BCWP}{BCWS} \quad (7.5)$$

- The *estimate at completion (EAC)* is given by the ACWP plus the difference between the BAC and the BCWP where the difference is divided by the CPI. In each case it is assumed that the cost efficiency will continue.

$$EAC = ACWP + \frac{BAC - BCWP}{CPI} \quad (7.6)$$

Table 7.7 NASA a priori risk classifications, from [18]

Characterization	Class A	Class B	Class C	Class D
Priority (Criticality to agency strategic plan) and acceptable risk level	High priority, Very low (minimized) risk	High priority, low risk	Medium priority, medium risk	Low priority, high risk
National significance	Very high	High	Medium	Low to medium
Complexity	Very high to high	High to medium	Medium to low	Medium to low
Mission lifetime (Primary Baseline Mission)	Long, >5 years	Medium, 2–5 years	Short, <2 years	Short, <2 years
Cost	High	High to medium	Medium to low	Low
Launch constraints	Critical	Medium	Few	Few to none
In-flight maintenance	N/A	Not feasible or difficult	Maybe feasible	Maybe feasible and planned
Alternative research opportunities or re-flight Opportunities	No Alternative or re-flight opportunities	Few or no alternative or re-flight opportunities	Some or few alternative or re-flight opportunities	Significant alternative or re-flight opportunities
Achievement of mission success criteria	All practical ara taken to achieve minimum risk to mission success. The highest assurance standards are used	Stringent assurance standards with only minor compromises in application to maintain a low risk to mission success	Medium risk of not achieving mission success may be acceptable. Reduced assurance standards are permitted	Medium or significant risk of not achieving mission success is permitted. Minimal assurance standards are permitted
Examples	HST, Cassini, JIMO, JWST	MER, MRO.Discovery payloads, ISS facility class payloads, ISS attached payloads	ESSP, Explorer payloads, MIDEX, ISS complex subtrack payloads	SPARTEN, GAS Can, technology demonstrators, simple ISS, express middeck and sub rack payloads, SMEX

- The estimate of schedule at completion (ESAC) is given by the SAC divided by the SPI, which assumes the same schedule efficiency will continue.

$$ESAC = \frac{SAC}{SPI} \quad (7.7)$$

An example of the status of a hypothetical project is given in Fig. 7.16.

There are three general techniques used to estimate costs: analogous, engineering build-up, and parametric. An analogous or analog estimate is based on the cost data from similar projects in which the differences due to inflation and technical differences can be estimated. The precision of this approach depends on the accuracy and detail with which the costs were tabulated, the accuracy with which the technical details are known, and the ability to extrapolate differences

of the new undertaking. The engineering build-up or bottom-up estimate is based on estimates from the lead engineers for each subsystem. This works well if the lead engineers have considerable experience with the products to be developed, but there is a tendency to include a margin at each level of management. This is the most costly and time consuming of the three methods, but generally the most accurate if the excess margins can be identified and negotiated. The parametric estimate is a top down process and uses computational models of similar developments to estimate the costs. These models use relationships between technical and program characteristics and costs based on prior developments. This approach depends on characterizing the product in terms of the model parameters and the similarity of the model developmental inputs to the current product. This can be a quick and less costly approach if the appropriate models are available.

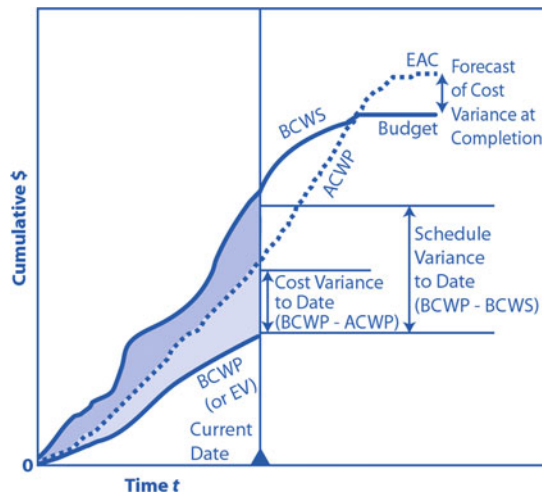


Fig. 7.16 Cost and schedule variances. Image: NASA/SP-2007-6105 [4]

7.10 Summary

Systems engineering is an interdisciplinary, structured procedure that guides the development of complex systems capable of satisfying the user's needs while balancing performance, cost, and schedule. Systems engineering is especially useful when there is no single obvious correct solution to satisfy the need, when the project is technically complex, or the management is distributed. This chapter described the fundamental concepts of systems engineering, the developmental life cycle, the development and flow-down of requirements, how to carry out technical assessments, the characteristics of the requisite organization, risk identification and management, and techniques to assess cost and schedule. Sufficient references are given to supporting documentation, especially from ESA and NASA.

References

- Schlager, J., "Systems Engineering: Key to Modern Development", *IRE Transactions EM-3*, doi:10.1109/IRET-EM.1956, pp. 64–66.
- Hall A. D., "A Methodology for Systems Engineering," Van Nostrand Reinhold, 1962.
- ECSS, "Space Engineering, System Engineering General Requirements," ECSS-E-ST-10C, ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, March 2009.
- NASA, "NASA Systems Engineering Handbook," NASA/SP-2007-6105 Rev. 1, NASA Headquarters, Washington, D.C., December 2007.
- Gruhl, W., "Lessons Learned, Cost/Schedule Assessment Guide," Internal presentation, NASA Comptroller's Office, 1992.
- ISO, Systems and Software Engineering – System Life Cycle Processes, ISO/IEC-15288:2008, Geneva, Switzerland, 2008.
- ISO, Systems and Software Engineering – Software Life Cycle Processes, ISO/IEC 12207:2008, Geneva, Switzerland, 2008.
- INCOSE, "Systems Engineering Handbook, v 3.2," 10 May 2010.
- ECSS, "System Engineering Guidelines," ECSS-E-HB-10, ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, in preparation.
- NASA, "NASA Systems Engineering Processes and Requirements w/Change 1," NPR 7123.1A, NASA Headquarters, Washington, D.C., March 2007.
- ECSS, "Space Project Management, Configuration and Information Management", ECSS-M-ST-40C, ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, 6 March 2009.
- NASA, "NASA Research and Technology Program and Project Management Requirements," NASA Headquarters, Washington, D.C., NPR 7120.8, 2008.
- ECSS, "Systems Engineering – Software", ECSS-E-ST-40C, ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, 6 March 2009.
- ECSS, "Software Product Assurance". Space Product Assurance," ECSS-Q-ST-80C, ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, 6 March 2009.
- NASA, "NASA Software Engineering Requirements," NPR 7150.2A, NASA Headquarters, Washington, D.C., 19 November 2009.
- ECSS, "Project Organization, ECSS-M-20B," ESA-ESTEC Requirements & Standards Division, Noordwijk, The Netherlands, June 2003.
- NASA, "NASA Space Flight Program and Project Management Requirements," NPR 7120.5E, NASA Headquarters, Washington, D.C., in preparation.
- NASA, Risk Classification for NASA Payloads, NPR 8705.4, NASA Headquarters, Washington, D.C., 2004.
- Blanchard, B. S., and Fabrycky, W. J., "Systems Engineering and Analysis," 4th Edition, Prentice Hall, Upper Saddle River, NJ, 2005.
- Griffin, M. D., and French, J. R., "Space Vehicle Design, 2nd Ed., AIAA Education Series, Reston, VA, 2004.
- Kossiakoff, A., and Sweet, W. N., "Systems Engineering Principles and Practice," John Wiley & Sons, New York, NY, 2003.
- Larson, W., Kirkpatrick, D., Sellers J., Thomas, L., and Verma, D., "Applied Space Systems Engineering," Space Technology Series, 2009.
- Pisacane, V. L., "Fundamentals of Space Systems," Oxford University Press, New York, N.Y., 2005.
- Sage, A. P., (Ed.) "Systems Engineering," John Wiley & Sons, New York, NY, 2010.

Further Reading

Christophe Bonnal, Alessandro Ciucci, Michael H. Obersteiner
and Oskar Haidn

Access to space requires the delivery of, typically, between 34 MJ/kg (low Earth orbit, LEO) and 58 MJ/kg (geostationary orbit, GEO) which, when considering the mass of our heaviest satellites, is a significant challenge. By comparison, a car driving 180 km/h has only the energy equivalent to 0.004 % (1,250 J/kg) of that required to access space.

To access space, a dedicated tool is required: a launcher or launch vehicle. Often also called a rocket, by assimilation of its propulsion principle, or a booster, as you need to boost your payload in order to have it gain the proper energy, but the most frequently used term is the word launcher: the machine that launches something into orbit.

The design of a launcher follows a succession of intricate activities known as the ‘system design loop’. A simplified example of such a loop is given in Fig. 8.1. These activities begin with the selection of the mission to be performed, and the associated constraints; for instance, the main mission of a new launcher may be the injection of a 4-ton satellite into a Sun-synchronous orbit at 800 km altitude, the associated constraints being the dimensions of the payload, its acceptable dynamic environment, and so on. This dictates the performance to be achieved by the launch vehicle.

The first technical activity of the system design loop is the choice of the number of propulsive stages, and their general

characteristics; size, diameter, type of propellants, and so on. From this can be deduced more precise propulsive characteristics, such as the engine cycle, the level of thrust, the size of the nozzles, and the general aerodynamic characteristics of the launcher; its drag, lift and moment coefficients throughout the complete flight regime in the atmosphere. Thereafter, the first trajectories can be computed, giving precious dimensioning data, mainly acceleration profile versus flight time, as well as dynamic pressure, enabling a first assessment of the induced environment.

Generation of the first trajectories enables assessment of the controllability of the launcher. This, in turn, gives the inputs for the engine tilting or fin steering, which leads to the determination of the general loads applied to the various parts of the launcher, the stresses induced to the structures by the inertia loads, and the dynamic pressure effects when the vehicle flies with a non-null angle of attack. This induced environment then facilitates the first general structural dimensioning of the launcher, i.e. selection of the material and thicknesses of each of the structural parts of the launcher. The controllability must be reassessed once the dynamical behavior of the launcher is known, as this has a strong influence on the control laws. Finally, a first mass breakdown of the launcher can be established, and a more effective performance estimate determined before once again starting the system design loop.

The system design loops are important, as there is no way to dimension a launcher without going (as the minimum) through all these steps. Furthermore, it is important to understand how deeply intricate these activities can be: if, for instance, an engine thrust is changed, then all the other design activities will be impacted, trajectories, accelerations, induced environment, control, mass, and so on. Tools such as Concurrent Design Facilities enable the complete design loop to be performed in near-real time, and allow rapid design trades.

The following paragraphs describe in a summarized manner the steps required to define and dimension a

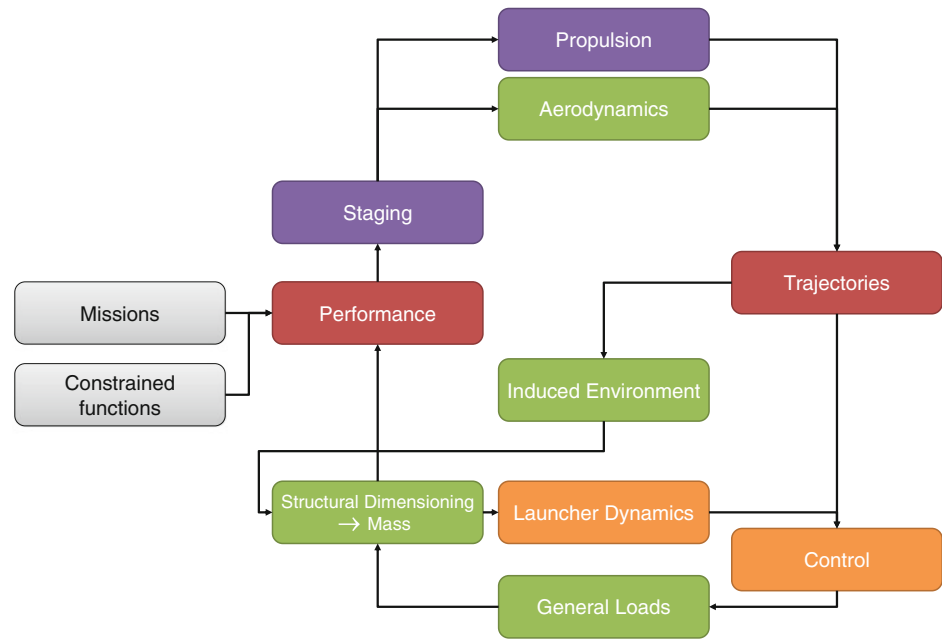
C. Bonnal (✉)
Directorate of Launchers, Centre National d’Études Spatiales
(CNES), Paris, France
e-mail: christophe.bonnal@cnes.fr

A. Ciucci
Headquarters, European Space Agency (ESA), Paris, France

M. H. Obersteiner
Competence Management, Airbus Defence & Space, Bremen,
Germany

O. Haidn
Institute of Flight Propulsion, Technische Universität München,
Garching, Germany

Fig. 8.1 Simplified system design loop



launcher, reviewing the constraints and solutions adopted for each associate technical topic.

8.1 Missions, Market, Functions and Constraints

The main function of a launch vehicle is to place an object, defined as the payload, into the specified orbit state. The target ‘state’ can be defined in many ways, as discussed in Chap. 4, using different but equivalent forms that are called either a ‘state vector’, when comprised solely of scalar magnitude terms, usually three position and three velocity components. Alternatively, an ‘element set’, when comprised of a set of geometric parameters, usually a mix of scalar magnitudes and angular representations. The technical criteria associated with this main function are the range of achievable orbits and payloads, and the quality of the final orbit, determined in terms of the velocity change that is needed to reach the final orbit. The target orbit state can be any of the orbit classifications discussed in Sect. 4.4, or beyond into interplanetary space.

The secondary functions are associated with the handling of the payload: the launcher must cope with its volume and mechanical interfaces, it must be protected from natural and induced environment, on-ground and in-flight, and it must be capable of data exchanges with the ground to enable its pre-launch operations and in-flight monitoring. The same is required for the launcher stages: environment control, data handling, power, air-conditioning, and so on.

Three sets of constraints can be identified. Product assurance requirements are the conditions to success

- Reliability corresponds to the contract with the payload, as it is the probability to achieve the required orbit.
- Availability is the probability to effectively launch when desired; it integrates notions such as mean time between failures (MTBF), and has a strong impact on the redundancy philosophy of the launcher layout, mainly regarding its electrical equipment.
- Maintainability is the probability to restore a flight status once an anomaly is detected; it integrates the notion of mean time to repair (MTTR), and leads to stringent constraints on the layout of the launcher and the accessibility of its most critical equipment.
- Safety requirements express the risk to populations by launch operations. It covers the launch site safety, including the integration operations, as well as the far range safety, which includes both nominal fall-out of the spent stages as well as the effects of anomalous events during flight (e.g. explosion or loss of propulsion or guidance).

Development constraints express the frame in which the development is carried out and cover constraints such as, in Europe specifically, an activities geographical return or *juste retour*, at country, region or industrial level. It also includes the programmatic constraints, such as the total development cost envelope and the time frame for the first flight. In some cases, the development constraints include some imposed solutions, such as when a given existing engine is to be considered for one of the stages, and also the growth potential requirements.

Production constraints are expressed by the operator: they can take the form of the launch frequency, the maximum mission cost, and the exchangeability between payloads.

8.2 Introduction to Launch, Notions of Staging

8.2.1 Summary of Applied Forces

A launcher is subject to a number of forces, the result of which creates the movement along the chosen trajectory. The principal applied force is the thrust generated. The thrust, or propulsive force F_P (N), is due to the ejection of gases in the direction opposite to the motion, with a mass flow \dot{m} (kg/s) and a velocity v_e (m/s). During the atmospheric phase, propulsive losses are associated to the area of the engine nozzle A_e (m²) on which the atmospheric pressure P_a (Pa) acts in the opposite direction to the thrust. That is

$$F_P = \dot{m} \cdot v_e - A_e \cdot P_a. \quad (8.1)$$

As introduced in Sect. 4.5.3, and will be discussed further in Chap. 11, the notion of specific impulse, I_{sp} , is used to characterize the efficiency of an engine, with the relation

$$v_e = g \cdot I_{sp} \quad (8.2)$$

with g standard acceleration due to gravity at sea-level, $g = 9.81$ m/s². Note that this equation is the same as Eq. 11.3 and related to Eq. 4.146. Recall that the I_{sp} is expressed in seconds. The classic thrust expression derived from Eqs. 8.1 and 8.2 therefore becomes

$$F_P = \dot{m} \cdot g \cdot I_{sp} - A_e \cdot P_a. \quad (8.3)$$

The direction of the propulsive force follows the axis of the nozzle and is applied at the level of the pivot point of the engine (if any). In the case of multiple chambers or nozzles, the propulsive force direction is determined by a combination of the n vectors (direction, intensity).

The second main force acting on a launcher is the aerodynamic force, which depends on the dynamic pressure that is generated by the relative velocity of the launcher with respect to the ambient atmosphere. That is

$$P_d = \frac{1}{2} \cdot \rho \cdot V_R^2 \quad (8.4)$$

with P_d (N), ρ (kg/m³) and V_R (m/s). The aerodynamic force is applied to the point called center of lift, which is where the integral of the elementary forces applied on the external surface of the launcher can be simplified into a perfect torque. As on an airplane, the aerodynamic force can be subdivided into a drag, or axial force, R_A , along the axis of the launcher and opposed to its movement, and a lift, or normal force, R_N , perpendicular to the launcher axis; see also Fig. 8.6. Other references are often used, defining drag and lift in the velocity frame instead of the launcher frame. The expressions of the drag R_A (N) and lift R_N (N) are given in Eq. 4.105 and again herein in a slightly different notation as

$$R_A = P_d \cdot S_{ref} \cdot C_A = \frac{1}{2} \cdot \rho \cdot v_R^2 \cdot S_{ref} \cdot C_A \quad (8.5)$$

and

$$R_N = P_d \cdot S_{ref} \cdot C_N = \frac{1}{2} \cdot \rho \cdot v_R^2 \cdot S_{ref} \cdot C_N \quad (8.6)$$

where C_A and C_N are respectively the drag and lift coefficients, and S_{ref} (m²), called the reference surface, is a normalization coefficient associated to the definition of the aerodynamic coefficients.

The third force acting on the launcher is its weight, applied at the center of gravity of the launcher, which varies with time as propellant is used. Finally, the wind is generally considered as a horizontal disturbance that modifies the relative velocity of the launcher and its incidence. The forces acting on a launcher are summarized in Fig. 8.2.

8.2.2 Equation of Dynamics

The forces defined on a launcher can be used to determine the trajectory and the stability of the launcher. The description given here is slightly simplified, considering movement in the vertical-horizontal plane, and force application points on the axis of the launcher, which generally is not the case.

Taking the X -axis as the main axis of the launcher and the Y -axis perpendicular in the vertical-horizontal plane, the projection of the forces defined in the previous section on the X and Y axes are

$$F_P \cdot \cos(\beta) - m \cdot g \cdot \sin(\theta) - R_A = m \cdot \Gamma_X \quad (8.7)$$

and

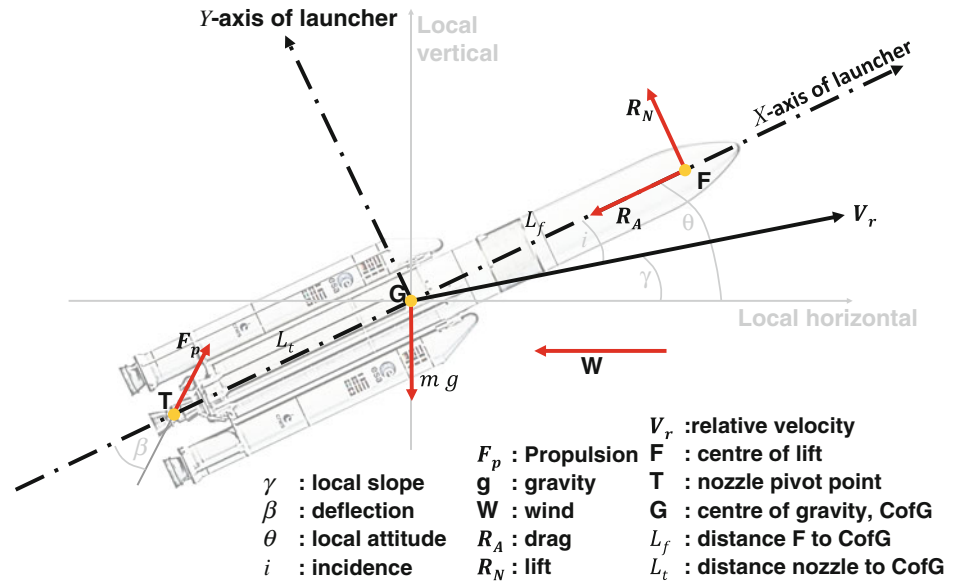
$$F_P \cdot \sin(\beta) - m \cdot g \cdot \cos(\theta) + R_N = m \cdot \Gamma_Y \quad (8.8)$$

where Γ_X and Γ_Y (m/s²) are the components of the acceleration of the launcher along the X - and Y -axes. Their time integral once, or twice, leads to the determination of the velocity profile, and of the trajectory of the launcher versus time. The evolution of this approach to three-dimensional trajectories with an out-of-plane component is straightforward, introducing the third dimension Z in the equations together with one additional attitude angle.

8.2.3 Stage Calculation, Basic Equations of Staging

When considering the equations of dynamics defined in Eq. 8.7 for the ideal case with no thrust deflection, no aerodynamics forces, and no angle of attack, the propulsive

Fig. 8.2 Forces acting on a launcher



acceleration Γ_P can simply be expressed as the ratio between the propulsive force F_P and the mass M of the launcher; these parameters are all time dependent.

If a given launcher stage functions in the time interval defined between i (initial) and f (final), the velocity increase generated by the propulsion during this interval can be written as

$$\Delta V_P = \int_i^f \Gamma_P \cdot dt = \int_i^f \frac{F_P}{m} \cdot dt = \int_i^f \frac{\dot{m} \cdot g \cdot I_{sp}}{m} \cdot dt. \quad (8.9)$$

It shall then be noted that the mass of the stage varies in time as

$$m = m_0 - \dot{m} \cdot (t_f - t_i) \quad (8.10)$$

with the mass flow \dot{m} defined as

$$\dot{m} = -\frac{dm}{dt}. \quad (8.11)$$

Applying Eqs. 8.10 and 8.11 into Eq. 8.9 gives

$$\Delta V_P = -\int_i^f g I_{sp} \frac{dm}{m} = g I_{sp} \ln\left(\frac{m_i}{m_f}\right) \quad (8.12)$$

with m_i and m_f the initial and the final masses of the stage (or launcher) at the beginning and end of the propulsion phase, respectively, and with \ln the neperian, or natural logarithm. Equation 8.12 is considered the fundamental equation of astronautics and was previously introduced in Eq. 4.146; it is known as the Tsiolkovsky, or ideal rocket equation because it was Tsiolkovsky who derived the modern form of the equation in 1903. However, as

discussed in Chap. 1 perhaps the earliest example of this kind of equation dates as far back as 1,813.

It is noted that the rocket equation represents the ideal case, with only one propulsive force acting along the axis of the launcher. In reality, a number of losses associated with the other forces acting on the launcher must be considered, in particular when the propulsive force is not applied along the proper direction, or when the angle of attack or incidence is not null. Equation 8.7 can be transformed into

$$\frac{dV}{dt} = \frac{F_p}{m} \cos(i + \beta) - \frac{R_A}{m} \cos i - \frac{R_N}{m} \sin i - g \cdot \sin \gamma \quad (8.13)$$

which, once time integrated leads to

$$\int_i^f \frac{dV}{dt} dt = \int_i^f \frac{F_p}{m} dt - \int_i^f \frac{F_p}{m} (1 - \cos(i + \beta)) dt - \int_i^f \frac{R_A}{m} \cos i \cdot dt - \int_i^f \frac{R_N}{m} \sin i \cdot dt - \int_i^f g \cdot \sin \gamma \cdot dt. \quad (8.14)$$

The various terms on the right are respectively the propulsive ΔV , the incidence and thrust orientation loss, the drag loss, the lift loss, and the gravity loss. The lift losses are always weak or null, thrust orientation and drag losses are important mainly for the first stage. To obtain an order of magnitude of these losses, take as an example the Ariane 5 ECA launcher in a geostationary transfer orbit (GTO) mission as detailed in Table 8.1. The losses can represent 10–25 % of the propulsive ΔV (or even more), so for an early design of a new launch vehicle, the order of magnitude of the losses must be considered.

Table 8.1 Typical losses acting on an Ariane 5 ECA GTO mission (CNES)

Losses	Incidence and deflection (m/s)	Drag (m/s)	Lift	Gravity (m/s)
Order of magnitude A5EC GTO	710	160	Weak or null	1,260

The mission ΔV is defined by the ΔV increase that is required for a given mission, the available propulsive ΔV_P , and the losses, i.e.

$$V_f - V_i = \Delta V_P - \text{losses}. \quad (8.15)$$

The final velocity, V_f , is also called the mission characteristic velocity and depends on the mission that is to be performed. Typical values are 7,500 m/s (LEO), 10,000 m/s (GTO) or more than 11,200 m/s (escape missions).

The initial velocity depends on the location of the launch pad and on the launch azimuth. That is

$$V_i = \Omega R_E \cos(\text{latitude}) \sin(\text{azimuth}) \quad (8.16)$$

where Ω represents the Earth's rotation rate and R_E the Earth's radius. When launching towards the geostationary orbit, using launch pads with low latitudes provides a benefit from the Earth's rotation induced velocity. Launching from the Sea Launch Odyssey platform, located at 0° latitude in the middle of the Pacific, gives an initial velocity of 464 m/s, while launching from the Guiana Space Center located at 5.2° is almost equivalent, with a gain of 463 m/s. Meanwhile, launching from Cape Canaveral located at 28.5° leads to a gain of 408 m/s, whereas launching from Baikonur, with a latitude of 45.6° but a launch azimuth constrained by the presence of China to a final inclination of 51.6° , leads to a gain limited to 288 m/s. Aiming at polar orbits, Sun-synchronous orbits, or highly inclined orbits change this initial velocity, reversing the sensitivity to latitude.

The initial velocity may also benefit from assistance at launch, for instance considering an air-launch, the launcher benefiting from an airplane as an initial stage. This is the case for instance with the Pegasus launcher, carried under a L1011 Tristar, benefiting from more than 200 m/s initial velocity. In the case of a launch below a balloon, the initial velocity is not changed, but the gravity losses are reduced by virtue of the release altitude.

8.2.4 Number of Stages

The ideal rocket equation, Eq. 8.12, can be generalized to multi-stage launchers. The case of a single stage launcher is easy to express, considering a dry mass, m_D , a propellant mass, m_P , a payload mass, m_{PL} and an average specific impulse I_{sp} , the available propulsive ΔV_P can be determined as

$$\Delta V_P = g \cdot I_{sp} \cdot \text{Ln} \left(\frac{m_D + m_P + m_{PL}}{m_D + m_{PL}} \right). \quad (8.17)$$

Considering the extreme case where no payload is launched, $m_{PL} = 0$, and introducing $k = \frac{m_D}{m_P}$, the structural ratio, the maximal available ΔV_P can be determined from Eq. 8.17 as

$$\Delta V_P = g \cdot I_{sp} \cdot \text{Ln} \frac{1+k}{k}. \quad (8.18)$$

This equation enables the theoretical analysis of a single-stage to orbit (SSTO) vehicle. Figure 8.3 shows the evolutions of ΔV_P as a function of the average specific impulse along the trajectory, and the structural ratio for such a SSTO. The figure shows that the assumptions required to reach orbit, with no payload, for a SSTO rocket propelled launcher are not realistic because even for an optimistic average specific impulse of 400 s it would require a structural ratio of 12 % to reach the lowest altitude Earth orbit. Comparably, considering a realistic overall structural ratio of 20 %, it would require an average specific impulse in the range of 500 s (average specific impulse means here the average along the complete trajectory, i.e. including all the atmospheric losses). It is not credible today to consider a rocket propelled SSTO launcher, however, propulsion based on air-breathing cycles can be much more efficient and could possible enable such a vehicle. Launchers are therefore, composed of several stages, practically two to four.

The calculation of a multi-stage launcher is easy, directly derived from the ideal rocket equation, considering one stage after the other, starting from the upper one, with each stage having a dry mass M_D equal to the sum of the dry mass of that stage added to the total mass of the stages above it, including payload and fairing.

8.2.5 Ballistic Phases

Depending on the mission definition, ballistic phases may have to be introduced to the launch profile. This is particularly the case when the launch pad latitude is high, restricting access to low inclination orbits, or when the final altitude is too high to be achieved by a single propelled phase, or simply to optimize the performance. During these ballistic phases, a number of perturbing forces may act on the vehicle, mainly those due to the residual atmosphere as discussed in Sect. 4.3.3, and must be taken into account. Typical examples are GTO launches from Cape Canaveral or Baikonur, or missions aiming at medium Earth orbit altitudes.

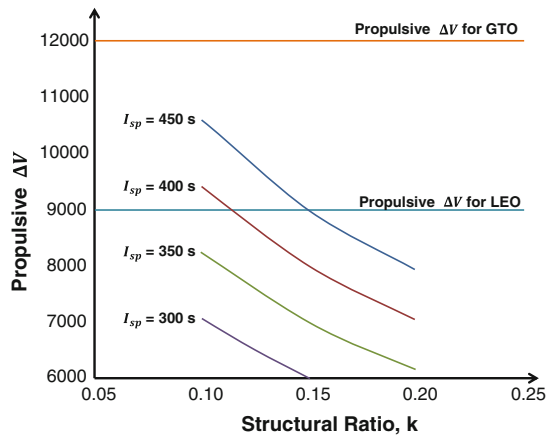


Fig. 8.3 Propulsive ΔV_p for a SSTO launcher

8.3 Launch Trajectory

The trajectory of a launcher is defined as the evolution over time of the position of its center of gravity. It can be defined with respect to a large number of different reference points, such as the launch pad, or a Keplerian reference frame in absolute Earth coordinates at the time of launch, leading to different sets of equations. Changing reference frame changes the equations of motion, but does not change the basic principles of computation.

8.3.1 Trajectory Design Constraints

Integrating Eqs. 8.7 and 8.8 is not very difficult *per se*, the difficulty resides in the fulfillment of all the constraints that must be applied.

Launcher acceleration: the maximum axial acceleration of a launcher is usually constrained by its payload. Typical maximal values are in the range of 45 m/s^2 for an automated payload, or 35 m/s^2 when the launcher is crewed. This constraint directly impacts the propulsion definition, and is also linked to the relative velocity through the drag term: i.e. for a given propulsion level, it is better to aim vertically in order to decrease the relative velocity.

Maximum dynamic pressure: the dynamic pressure is used to dimension the structures of the launcher, stages, inter-stages structures, and fairing. Both the dynamic pressure and its product with the angle of attack lead to stresses on the structures and on the control requirements. The worst part of the flight is the so-called ‘maximum dynamic pressure’, which usually occurs close to 70 s into the flight. This imposes a constraint to not fly too low for a given velocity.

Maximal thermal constraint: the structures are dimensioned by the thermal fluxes they encounter during the atmospheric phase, which are proportional to the cube of the

relative velocity (depending on the flight phase, it may vary as $V_R^{3.15}$). This imposes a constraint to not fly too low for a given relative velocity as the atmospheric density decreases rapidly with altitude.

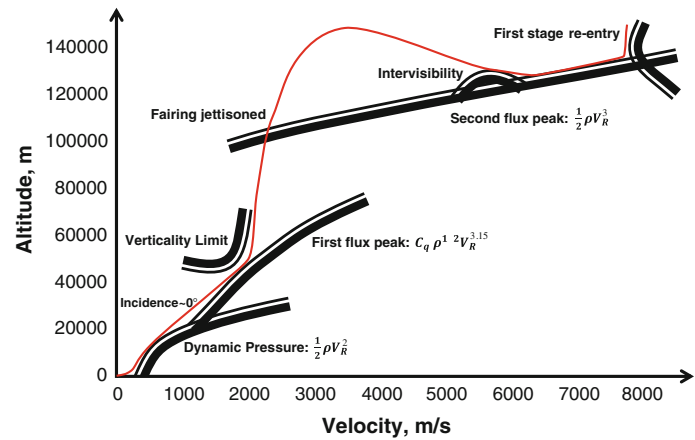
Pitch-over: a launcher usually starts vertically on its launch pad (it may be slightly tilted for smaller launchers); however, at the injection into orbit, the velocity vector is usually more or less along the local horizontal. This means the velocity vector must be progressively tilted from vertical to horizontal. This pitch-over maneuver is usually triggered shortly after the launch, typically after some 10 s of flight, and the launcher then follows a naturally curved trajectory called a gravity turn, defined by a zero angle of attack, at least until it gets out of the atmosphere.

The trajectory must also pass through some gaps between visibility regions from the ground stations; consequently the trajectory is constrained to impose a higher altitude than optimal in some of the regions, depending on the overlap time.

Safety constraints are typically very stringent and can strongly influence the trajectory. The first safety constraint is the near-field constraint. All around the launch pad up to distances of tens of kilometers, there will be installations or even cities that have to be protected during a launch. As a general assumption, consider that the launcher may explode at any instant of its trajectory, spreading dangerous debris or pollution over a wide area. The size of this area depends on the altitude and the velocity vector at the instant of explosion, together with the explosion model (which itself is a function of the type of propulsion and the quantity of propellant involved). To that extent, two limits are defined, one on the vertical, and the other as a zone on the horizontal plane matching the ground track. This leads to constraints on the rate of the pitch-over and on the launch azimuth, taking into account the wind.

The second safety constraint is the long-range safety, addressing dropped stages or debris in both nominal and abnormal cases. This constraint is usually expressed in the legal framework associated with the operations of the launch system. The nominal descent of used stages must take place away from any landmass or islands with a given probability. For instance, this can require that the debris zone shall be in international waters with a probability better than 1×10^{-4} . In the case of an abnormal event occurring during the ascent, such as loss of propulsion or an explosion, it is impossible to guarantee that debris will not fall on land, or even on inhabited zones; the constraint is then expressed as the minimization of the casualty risk associated with the launch, which must be below a specified threshold. This risk is computed as the time integral over the mission of the product of the probability of losing propulsion or exploding, multiplied by the conditional probability of generating casualties at that location.

Fig. 8.4 Trajectory constraints—Ariane 5 GTO mission; nominal trajectory in red



These constraints are often difficult to integrate correctly, as some of them may be antagonist. For instance, it may be best to launch as vertical as possible in order to lower acceleration and dynamic pressure, but as horizontal as possible at the same time to appease the near-field safety constraint.

Figure 8.4 illustrates the effect of some of the trajectory constraints mentioned for the case of an Ariane 5 GTO mission. It presents the nominal trajectory in red, in the plane of velocity (m/s) versus altitude (m) for the first phases of the mission. A compromise between structural dimensioning (dynamic pressure and heat flux) and safety constraints (verticality limits) makes the flight corridor tight.

8.3.2 Trajectory Optimization

Trajectory optimization is similar to any mathematical optimization and follows the same expression: a state vector is defined, composed of the vector position and velocity of the launcher center of gravity, of its mass, and of the set of parameters on which the optimization can be performed, including payload mass and combustion time. The state vector is sought to describe the thrust magnitude and direction, expressed as two angles.

Intermediate constraints can be defined, such as maximum dynamic pressure and acceleration, heat flux, minimum inter-visibility between ground stations, reentry trajectories for dropped stages, and so forth. The performance criterion is thereafter defined, which can vary strongly depending on the case under consideration.

During the early design of a launcher, the optimization may be done considering a given objective payload for the reference mission, seeking to minimize the total mass at liftoff. Then, once the design is frozen, trajectory optimization will be done for other missions, maximizing the payload mass.

During the production phase, once both the launcher and the payload are defined, the optimization process can

maximize other parameters, such as safety on the ground, ground stations visibility or performance reserve.

The optimization itself is classical, for which various techniques exist such as direct methods for which variations around a local solution are computed, or indirect methods where the necessary conditions of optimality are satisfied before improving the criterion; the most commonly used indirect method follows the Pontryagin principle which was established in 1962.

8.3.3 Trajectory Simulations

Once a launcher trajectory is optimized, simulations can be performed considering variation around the nominal values within each input to a domain, representing either dispersions or unknowns. This will include a model of the real avionics performances, and can even include potential failures. Simulating a trajectory with these dispersed parameters enables the robustness of the nominal solution previously determined to be established, and to ascertain a flight domain. This process is fundamental, and is the only real evaluation of the probability of achieving the required orbit while complying with all the constraints. Simulations are usually performed using mathematical tools such as a Monte-Carlo analysis.

8.4 Launcher Guidance, Navigation and Control

The launch vehicle, after liftoff, is typically completely autonomous and must place the payload(s) in the proper orbit, regardless of any disturbances. To that extent, the vehicle must ensure some management functions: ignition, cut-off, separations (stages, fairing, and so forth), pressurization management, and flight control functions including guidance, navigation, and control (GNC), failure detection, redundancy, and so forth.

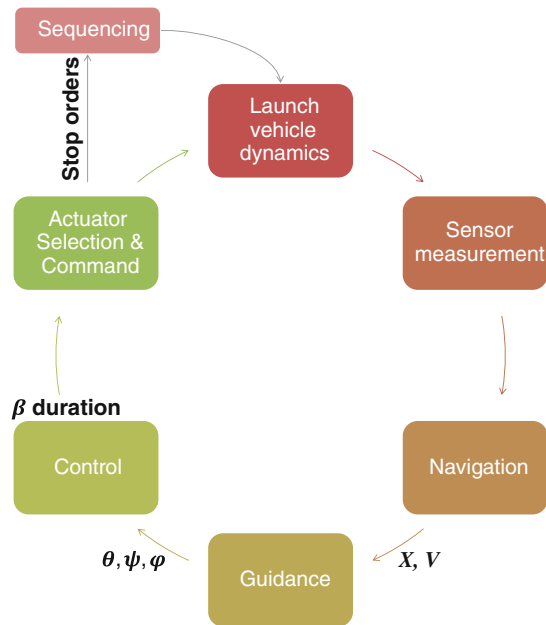


Fig. 8.5 Launcher guidance, navigation and control loop

Guidance, navigation, and control is discussed in detail in Chap. 12, but for convenience the GNC loop of a launcher is presented in Fig. 8.5. It enables the launcher to thrust in the proper direction. The aerodynamic stability of the launcher will often drive the frequency of the control loop, for instance on Ariane 5 this frequency is 14 Hz.

8.4.1 Launcher Navigation

The navigation function aims to give at any time during the flight the position and velocity of the launcher. It must account for the real shape of the Earth, and is typically based on single, then double integration of accelerometers during the propulsive phases. Navigation can be simplified during long ballistic phases by using an analytic integration of the trajectory.

Navigation can also make use of information coming from the global navigation satellite systems such as GPS, Galileo, BeiDou-Compass, or GLONASS, but the required mission precision generally does not impose the use of such service, and independent operation is often preferred.

Depending on the accuracy of the sensors, the orbital parameters are known with a given precision and inter-relation, generally expressed as a covariance matrix, as in Table 8.2.

8.4.2 Guidance

Once the effective position of the launcher is known, it is compared in the on-board computer to the theoretical

position of the launcher at a given time of flight computed during the final mission analysis, and the difference between the two positions is determined. The guidance function aims to provide a thrust direction, during propulsive phases, in order to follow the optimal trajectory while minimizing the propellant consumption and fulfilling the trajectory constraints (final orbit, thermal fluxes, visibility from telemetry stations, and so forth).

Guidance is a three-dimensional computation that takes into account only the position of the center of gravity of the launcher. Its outputs are the Euler angles defining the launcher's attitude. The atmospheric phase is performed following a table containing the pre-computed values of the pitch and yaw angles as a function of time, or of relative velocity, depending on the launcher and its robustness.

Beyond the atmosphere, the trajectory is subdivided into segments, the end of each corresponding to a key event of the launch, for example, separation of the fairing, jettisoning of the stages, inter-visibility between ground stations, and so forth. At the end of each segment, a number of parameters have to be verified, such as orbital parameters and flux constraints. In some specific cases, where for example a solid propelled stage is used, an additional constraint may be imposed to control the impact location of the empty stage on the ground. The 'neutral axis maneuver' defines a dedicated guidance law that is applied during the last seconds of propulsion in order to generate a pitch maneuver that will maintain a fixed impact point of the dropped stage by accounting for any propulsion dispersions previously encountered.

8.4.3 Control

The control function aims to maintain the behavior of the launcher around a given trajectory, with the dual goal to impart a zero-degree angle of attack during the flight in order to lower the mechanical constraints, and to guarantee the required injection accuracy in orbit.

Two sets of disturbances must be accounted for

- Internal disturbances come from the flexible modes of the structure, the fluid sloshing modes, and more generally all uncertainties or dispersions such as delays, measurement precisions, as so forth.
- External disturbances are mainly encountered during the atmospheric phase, such as the influence of wind.

The control loop has to deal with objectives of stability. That is, a launcher is generally naturally unstable, meaning that without control it will tilt and diverge from its nominal trajectory before breaking apart. This stability objective is mainly dealt with in the frequency domain, whilst the

Table 8.2 Example of a covariance matrix on orbital parameter in GTO (CNES)

	a (km)	e_x (-)	e_y (-)	i (rad)	Ω (rad)	$\omega + M$ (rad)
a (km)	4.89×10^{-0}	6.91×10^{-4}	-4.42×10^{-5}	-1.21×10^{-4}	-1.25×10^{-4}	-4.11×10^{-3}
e_x (-)	6.91×10^{-4}	8.52×10^{-8}	-1.16×10^{-8}	-1.49×10^{-8}	-8.74×10^{-9}	-5.60×10^{-7}
e_y (-)	-4.42×10^{-5}	-1.16×10^{-8}	4.48×10^{-8}	4.71×10^{-9}	1.18×10^{-8}	-5.05×10^{-8}
i (rad)	-1.21×10^{-4}	-1.49×10^{-8}	4.71×10^{-9}	2.79×10^{-7}	3.92×10^{-8}	7.10×10^{-8}
Ω (rad)	-1.25×10^{-4}	-8.74×10^{-9}	1.18×10^{-8}	3.92×10^{-8}	4.02×10^{-9}	2.05×10^{-8}
$\omega + M$ (rad)	-4.11×10^{-4}	-5.60×10^{-7}	5.05×10^{-8}	7.10×10^{-8}	2.05×10^{-8}	4.20×10^{-6}

control loop will also deal with reactivity constraints in the time domain.

Various sensors are used as inputs to the control loop, such as inertial systems giving angular measurements, and rate gyros giving angular velocities. Accelerometers also provide transverse acceleration data that can be used to limit the angle of attack during the atmospheric flight. The output of the control loop will command the various actuators that swivel the engines during the propulsive phases, or open the secondary thrusters during the ballistic phases.

The stability conditions of a launcher can be determined by considering the forces acting on a launcher, as seen in Fig. 8.2. The fundamental principle of dynamics gives the conditions of stability both in position and in angle

$$\sum \vec{F}_{forces} = \frac{d(m\vec{V})}{dt} \quad (8.19)$$

and

$$\sum \vec{T}_{torques} = \frac{d(I\vec{\omega})}{dt} \quad (8.20)$$

with m the mass of the launcher, V its vectorial velocity, I its inertia matrix, and ω its angular velocity as a vector. As seen previously, Eq. 8.19 enables the computation of the launcher trajectory. Equation 8.20 is used to determine the conditions of stability of the launcher around one axis, here chosen as the pitch axis

$$\ddot{\Theta} = \frac{P_d \cdot S_{ref} \cdot C_{mi} \cdot L_f}{I} \cdot i + \frac{F_p \cdot L_t}{I} \cdot \beta \quad (8.21)$$

where (using the same notations as in Fig. 8.2) Θ is the local attitude angle along the pitch axis, P_d is the dynamic pressure, S_{ref} is the reference surface, C_{mi} is the pitch coefficient derivial with respect to incidence, L_f is the distance from the center of lift to center of gravity, I is the inertia of the launcher along the pitch axis, i is the incidence, F_p is the propulsive force, L_t is the distance from engine gimbal to the center of gravity, and β is the deflection angle of the engine. This equation, fundamental for all the control analyses, is often written as

$$\ddot{\Theta} = A6 \cdot i + K1 \cdot \beta \quad (8.22)$$

with

$$A6 = \frac{P_d \cdot S_{ref} \cdot C_{mi} \cdot L_f}{I} \quad (8.23)$$

and

$$K1 = \frac{F_p \cdot L_t}{I} \quad (8.24)$$

$A6$ expresses the aerodynamic efficiency, or the propensity of the launcher to increase its incidence when subjected to dynamic pressure. A positive $A6$ denotes an unstable launcher, and a negative one a stable launcher. $K1$ expresses the deflection efficiency, or the propensity of the launcher to come back to a null incidence in response to the swiveling of its propulsion.

Using the Laplace formalism, the stability conditions can be written in a simplified way as an open-loop transfer functions between attitude and deflection

$$F(s) = \frac{\Theta}{\beta} = \frac{K1}{s^2 - A6} \quad (8.25)$$

which then enables the use of all the powerful tools from automatic control theory, the study of gain margin or phase margin, the introduction of perturbations, local modes, sloshing, and so forth.

Most large launchers are aerodynamically unstable. Solutions to this instability are diverse: fins can be added at the bottom of the launcher to improve the $A6$, or vernier thrusters can be used in addition to the gimbaling of the main engines to improve the $K1$.

8.5 Mechanical Conception

The mechanical dimensioning of a launcher follows the same methodology as that of a satellite, or even more generally of an airplane or any other vehicle. The major specificities come from the domain of flight, covering all of the velocity domain up to orbit, including transition through

the atmosphere, and also the peculiar shape of the stages of a launcher; generally made of long, very thin, shell like cylinders, metallic tanks, and metallic or composite intermediate structures.

The first step is to identify all the dimensioning cases associated with aerodynamics, thrust of the rocket engines, transitory phases such as liftoff or stage separations, shocks, thermal environment, as so forth. Then these dimensioning cases have to be transformed into mechanical constraints, traction or compression, applicable to all the elements constituting the structure of the launcher. Finally, material choices have to be made, and the different thicknesses of the elements defined taking into account manufacturing constraints.

8.5.1 External Environment

Aerology

The first element that has to be taken into account in the mechanical dimensioning of a launcher is the wind, both on the ground and in flight. Wind is a highly variable and turbulent phenomenon, which depends on the season and the altitude. It is composed of a mean component and fluctuant parts. In flight, wind plays a major role in the control of the launcher, as it adds some angle of attack during the atmospheric phase.

Knowledge of the wind profiles to be applied comes mainly from all the measurements performed by the meteorological team(s) of a launch base. The measures can be done using sounding rockets, meteorological balloons, or radio soundings. The merging of these measures, average winds established with balloons, and the meso-scale wind profile determined by radio-sounding, gives what is called 'real' winds. As an output, a statistical base is used to generate dimensioning winds and enable realistic statistical analyses, thus avoiding the need to take the worst case scenario at each altitude. Finally, some gusts may be added, following a profile leading to a maximum wind (typically 9 m/s) over a given length (for instance 100 m) determined by the length of the launcher.

Aerodynamics

Steady aerodynamics are fundamental for launch vehicle dimensioning as they play a major role in the computation of performances, through the drag, the control at liftoff, the structural dimensioning, and the aero-heating of the structures.

Two forces are encountered: pressure forces, normal to a structural element, and tangential friction forces. Practically, viscous effects are felt only on a limited zone very close to the surface, called the boundary layer, which marks the limit between computations in real fluid and viscous fluid. The representativeness of such phenomena between

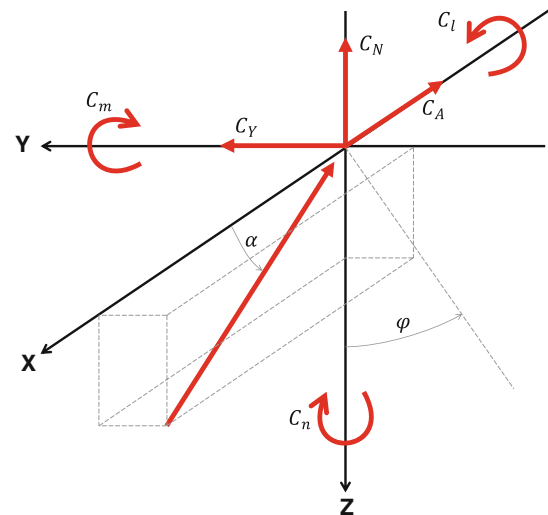


Fig. 8.6 Typical definition of global aerodynamic coefficients

ground tests and flight is determined by considering the Reynolds number, introduced in [Chap. 5](#) as ratio between convective and viscous phenomenon. The Reynolds number marks the limit between laminar and turbulent flows.

At general launcher level, six coefficients are defined: three force coefficients and three moments, respectively C_A (axial, or drag), C_N (normal, or lift), C_Y (side), C_l (roll), C_m (pitch), and C_n (yaw). Figure 8.6 gives a typical structure of these global aerodynamic coefficients. An additional global aerodynamic coefficient that is widely used is $C_{N\alpha}$, which is the linearized deprival of the lift versus the angle of attack; it is a key parameter in the definition of the aerodynamic efficiency $A6$ used in control stability studies.

Locally, on an elementary surface, two coefficients are defined: the pressure coefficient, which the ratio between the infinite pressure (i.e. sum of static and dynamic pressures) and the local pressure applied on the element, and the friction coefficient. These coefficients are then integrated all over the elementary surfaces of the launcher to obtain the distributed aerodynamic loads and these are then used for the mechanical dimensioning. Figure 8.7 gives a typical example for the Ariane 5 launcher in the yaw plane at Mach 0.98.

Some unsteady aerodynamic phenomenon should also be taken into account, such as the unsteady acoustic aerodynamic effects that generate random pressure fluctuations. Acoustic effects are generally not critical in dimensioning the main launcher structure, but are a significant contributor to large-scale vibrations (20–2,000 Hz). Such excitations are important for launcher equipment and payloads (satellite or human), and are most important during liftoff (noise from engine plume and blast wave), during transonic flight, and maximum dynamic pressure (buffeting), depending on the

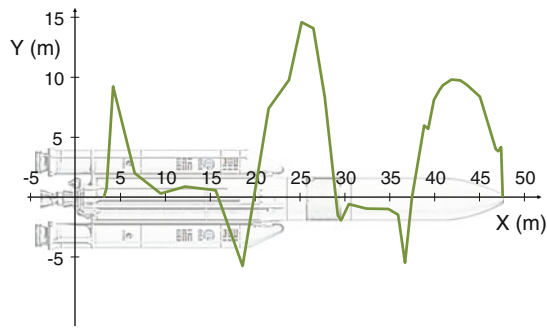


Fig. 8.7 Distributed aerodynamic loads; example of Ariane 5 in the yaw plane at Mach 0.98 (CNES)

coupled design of the launch pad, the launcher and the type of propulsion used.

The blast wave induced by the exhaust plume can be encountered on large launchers at liftoff, either coming directly from the duct inlets and raising along the launcher (ignition overpressure, IOP), or coming from the duct outlets and laterally exciting the launcher (duct overpressure, DOP). This deterministic overpressure field leads to high-level lateral excitation at low frequency, typically below 20 Hz.

Several other unsteady aerodynamic phenomenon can also play an important role in the structural dimensioning of a launcher, depending on its shape definition. One is buffeting, a pressure fluctuation imposed mainly on the fairing at liftoff. Another is base instability, which occurs when the cross-section of the launcher is sharply reduced, leading to a very turbulent flow recirculation, with potential reattachment of the flow on the nozzle of the engines inducing strong side loads and vibrations transmitted to the complete launcher.

Thermal dimensioning

A launcher is submitted to extreme thermal conditions

- Internal temperatures range from very cold, with cryogenic temperatures such as the 20 K of liquid hydrogen, to very hot, such as the 3,000 K within a rocket engine combustion chamber.
- The external conditions are generated by the very high speed in the atmosphere, leading to friction and compression of the air and an increase of temperature that imposes a need for dedicated thermal protection; mainly for a reentering vehicle.
- Outside the atmosphere, a launcher stage is highly sensitivity to radiation, which can lead to very hot structures facing the Sun, and to very cold temperatures in the shadow, therefore imposing a high thermal gradient.

There is a need to define properly the thermal protections to be applied on every element of a launch vehicle, taking

into account the various heat transfer mechanisms and the various phases of its life (including ground phases). For a cryogenic upper stage, its mission may require separate tanks to facilitate the necessary insulation and minimize the thermal inputs.

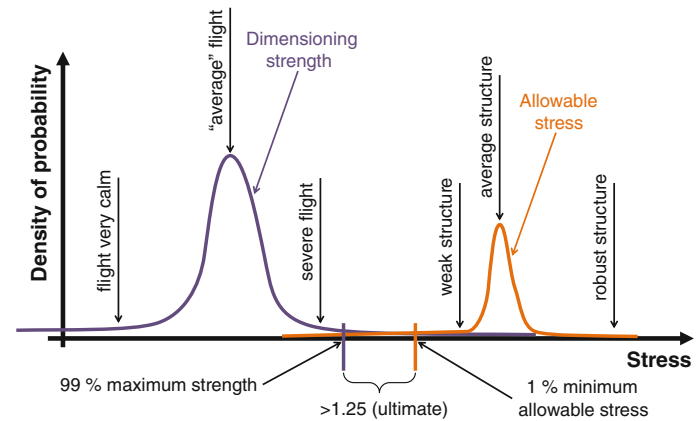
Three heat transfer mechanisms must be taken into account

- Convection is the transfer process executed by the flow of a fluid (liquid or gas medium); it is negligible outside the atmosphere. Convection can be natural, due to the natural exchange between walls and atmosphere, or forced, mainly in the cavities which must remain within a given temperature range during the flight. The estimation of convective fluxes is a difficult task, leading to the use of semi-empirical formulas and computational fluid dynamic (CFD) models validated by ground tests. It is globally proportional to the difference in temperature between air and structure.
- Conduction is the transfer process associated with the propagation of heat inside materials without any motion (solid medium). Materials range from good conductors, typically metals such as aluminum, to good isolators, typically foams used to protect cryogenic tanks. It is also proportional to the difference in temperature, following Fourier's law. It should be noted that for a reentry vehicle, due to the heat soak-back effect (inertia effect), the temperature of the structure may reach its maximum after reentry.
- Radiation is the transfer process linked to the energy carried out by an electromagnetic wave (no medium required), and is the main driver in vacuum. It is proportional to the temperature difference raised to the fourth power. The radiative exchange factor is usually computed using a Monte-Carlo method (randomization on ray directions), and takes into account wall geometry (areas, view factors), thermo-optical properties, and transfer by multiple reflections. A significant contributor to these thermal fluxes is the radiative heat flux induced by rocket engines and the rockets jet plumes in the infrared band.

The thermal dimensioning of a structure must take into account all three heat flux transfer processes, both on the ground and in flight, which leads to complex computations for the two dimensioning cases, hot and cold. A variety of control solutions are available, both passive (various thermal protections) and active (phase change material, forced circulation, and so forth).

8.5.2 General Loads

The general loads applied to the structure(s) of a launcher are the mechanical loads generated at the global launcher

Fig. 8.8 Warner diagram

scale, to which are added the local effects. The determination of the inputs for the structure dimensioning is usually performed using a statistical approach following a strength-stress theory schematized by a Warner diagram, as illustrated in Fig. 8.8. This diagram establishes firstly the distribution of loads, following the left curve. Depending on the dispersion of the stresses, for instance linked to the atmospheric density, the propulsion characteristics, or the effective trajectory followed, a Gaussian distribution can be defined with a maximum at the ‘average’ flight, and higher values for more ‘severe’ ones. This defines the dimensioning strength to which the structure will be subjected. Similarly, the distribution of strength can be defined following the effective robustness of the structure, on the right curve, depending for instance on the effective specific strength of the materials used. This in turn defines the allowable stress curve to which the structure can be subjected.

A safety factor, j , is defined as the ratio between the minimal allowable stress at 1 % probability, and the maximal strength at 99 % probability. Such a safety factor is usually considered at 1.25 for ultimate stresses.

The general loads computation is performed by considering ‘cuts’ in the launcher at various altitudes, called stations, where the loads induced to the lower part are computed (aerodynamics, thrust and inertial effects) while also accounting for the loads coming from the upper parts. Three components are then defined: N which is the normal load, T which is the shear load, and M the moment applied to the station. As the station is stationary at any time, due to structural resistance of the junction, the same three components apply one-for-one at both the lower and the upper parts of a given station.

General loads can be split into quasi-static loads, which represent the slow temporal evolution of the loads, and dynamic loads, which are the temporal evolutions of the loads for various bands of frequencies. This dynamic part can be analyzed either by identifying the stresses as a function of the time of their occurrence, or in the frequency domain, with each phenomenon being linked to one band of frequency.

8.5.3 Architecture and Structural Dimensioning

A launcher is usually composed of several propulsive stages linked together by inter-stage skirts, or struts and bolts (for the side boosters), has a vehicle equipment bay, and is topped by a fairing protecting the payload(s).

Each propulsive stage is itself composed of several propellant and pressurant tanks (if liquid), and engine(s) linked to the main structure through thrust frames. A stage’s overall architecture may be a function of several considerations, such as the effects of adjacent parameters, design for manufacture, geometrical and functional or even technical constraints, kinematics and dynamics at the system level, and any number of secondary requirements.

The propellant tanks can themselves present a wide variety of designs, with common bulkheads or separate tanks, or inclusive tanks with one nested inside the other. A diverse range of shapes can also be used: for example cylindrical-ellipsoidal, cylindrical-spherical, toroidal, and conical. Their link to the main structure depends on the overall architecture: they can be integral, carrying all the loads, clustered, or suspended inside the structure.

The selection of material(s) for the primary structure(s) is a complex function of the influence of the stage on launcher performance and its various trade-off criteria, a compromise between production and operational aspects, and a function of the specific strength, stiffness and required degree of anisotropy. Such structures are generally made of metal or composite. Metals are usually aluminum alloys including aluminum–lithium, steel, and titanium. Composites are often carbon fiber reinforced plastic (CFRP). Structures are also often based on sandwich technologies, for instance aluminum sandwich within two CFRP face-sheets. Launcher design generally a compromise between achieving a high strength level and a lightweight concept. The most influential parameters are strength, elasticity, density, fracture toughness, manufacturing, fuel compatibility, and corrosion resistance. Table 8.3 lists the main

Table 8.3 Main mechanical characteristics of the most common materials; ρ is the density of the material, σ_r its rupture strength, E its Young's modulus

Material	ρ (kg/m ³)	σ_r (kg/m ³)	E (M Pa)	$\frac{\sigma}{\rho}$ (10 ³)	$\frac{E}{\rho}$
Magnesium alloys	1,800	230	42,000	128	23
Titanium alloys	4,500	920	115,000	204	25
Aluminum alloys	2,800	400	72,000	143	26
Steel alloys	7,800	1,050	205,000	135	26
		1,800		231	
Kevlar 49 fiber/epoxy (unidirectional)	1,370	1,600	85,000	1,168	62
HR carbon fiber/epoxy (unidirectional)	1,560	1,400	130,000	897	83
HM carbon fiber/epoxy (unidirectional)	1,660	1,100	250,000	663	150

mechanical characteristics of common materials used for main structures and tanks.

The maximum expected operating pressure (MEOP), which often reaches several hundred bars, dimensions the high-pressure vessels to be used for pressurants. These tanks are usually made of titanium or stainless steel, overwrapped with carbon fiber layers.

Various load levels have to be considered for the structure dimensioning, each of which has with its own safety coefficient J

- *Limit load*—maximal load level in the structure life (at 99 %).
- Yield level (for metallic items)—limit load times J_e safety factor.
- Ultimate level—limit load times J_r safety factor.
- *Acceptance/proof*—limit load times J_p safety factor (proof).

Three different sizing methods are considered

- The strength method is applied to structures in traction, including pressure vessel walls.
- Stability is associated with the risk of a structure buckling under compression. This stress can be relieved by internal pressure inside the structure, and corresponding formulas are well known for isotropic structures; however, it is much more complex for anisotropic structures such as composites. Stiffeners may be added to the structure to improve its tolerance to buckling; various shapes of stiffeners are commonly used, straight, or Ω for instance.
- Stiffness sizing is associated with requirements at the stage or launch vehicle level, considering the frequency of the main structural modes. Stiffness is a function of the Young's modulus of the selected material.

These methods are classical, not specific to launch vehicles, and are widely used for any structural dimensioning in any mechanical domain.

8.6 Launch Vehicle Rocket Propulsion

The main function of a rocket propulsion system is to generate a force, called thrust, which when applied to the adjacent structure induces an acceleration in the direction opposite to the engine flow.

On a launch system, such engines can be used either to generate the main acceleration of the launcher, enabling it to reach the desired velocity, for instance the orbital velocity in the case of an orbital stage or to generate a torque on the stage following any of the six degrees of freedom in order to guarantee the attitude control of the assembly. This torque is composed of small forces along the three translational axes, or the three moments around the three axes to enable the proper orientation, for instance prior to the separation of the payloads.

The principle of a rocket engine is to eject rapidly gases that, by equal and opposite reaction, produce the thrust. Rocket engines can be based on a wide variety of propellants, which are discussed in detail in [Chap. 11](#), and are introduced here

- Cold gases can be used when only low thrusts are required, for instance for the attitude control of an upper stage prior to payload separation. The cold gases traditionally used for this can be gaseous hydrogen or nitrogen. There is no combustion, just a pressure decrease generating the ejected mass flow.
- Monopropellants are often used on small launchers for the propulsion of the orbital stage, or as attitude control systems for larger ones. The principle is to use a highly exothermic propellant decomposed on a catalytic bed. The heat released by the decomposition increases the enthalpy of the gases and the ejection velocity. The monopropellant most widely used today is hydrazine (N_2H_4).
- Bipropellant can generate thrust from several newtons up to 8,000+ kN for larger engines. The principle is to perform an oxido-reduction reaction between an oxidizer

and a fuel, with the chemical reaction occurring within the combustion chamber generating high enthalpy gases that are ejected through the nozzle of the engine. The oxidizers most commonly used are oxygen and N_2O_4 and derived components. The most frequent fuels are hydrogen, kerosene and products derived from hydrazine, such as monomethyl-hydrazine (MMH) for upper stages or dimethyl-hydrazine (UDMH), mixed or not with another fuel, for, mainly lower stage. Methane and liquid natural gas are considered for future engines.

- Solid propellant motors are also based on oxido-reduction reactions, but the oxidizer and the fuel are mixed, generally blended with a binder, and stored in a solid form inside the engine body itself.

Whatever the engine, the principle is always the same: a high-pressure chamber leads to a convergent-divergent throat connected to a nozzle that accelerates the gases to generate thrust.

Characteristic Velocity

Considering the flow rate equation at the nozzle throat

$$\dot{m} = A_t \cdot \rho \cdot V \quad (8.26)$$

with \dot{m} the mass flow (kg/s), A_t the throat area (m^2), ρ the density of the gases inside the chamber (kg/m^3), and V the exhaust velocity of gases at the throat plane (m/s), an important coefficient can be derived called the characteristic velocity, denoted c^* . This is representative of what happens in the combustion chamber

$$c^* = \frac{P_c \cdot S_t}{\dot{m}} = \left(\frac{\gamma + 1}{2}\right)^{\frac{\gamma+1}{2(\gamma-1)}} \cdot \sqrt{\frac{R \cdot T_c}{\gamma \cdot M}} \quad (8.27)$$

with P_c the combustion pressure (Pa), γ the adiabatic constant for the gases considered, R the perfect gases constant (J/mol/K), T_c the combustion temperature (K), and M the molar mass of the gases (kg/mol); c^* is in m/s.

The characteristic velocity depends only on the gas temperature and on the composition of the gases in the combustion chamber, so it is characteristic of the efficiency of the chemical reaction inside the combustion chamber. It enables a comparison between propellants (c^* is the reverse of C_d or discharge coefficient, the ratio of mass flow rate, \dot{m} , at the nozzle exhaust to that of an ideal nozzle expanding an identical fluid over the same pressures, i.e. a ratio of actual to theoretical discharge)

Thrust Coefficient

When combining the equation of propulsion, the flow rate at the nozzle throat, Eq. 8.26, and the velocity of the gases at the outlet of the nozzle, the ejection velocity, v_e , may be written, in (m/s), as

$$v_e = \sqrt{\frac{2\gamma}{\gamma-1} \frac{RT_c}{M} \left[1 - \left(\frac{P_e}{P_c}\right)^{\frac{\gamma-1}{\gamma}}\right]} \quad (8.28)$$

with P_e the pressure at the exit of the nozzle (Pa). An important coefficient can be determined, characterizing the efficiency of the acceleration of the gases inside the nozzle, called the thrust coefficient, C_F , that is

$$C_F = \sqrt{\frac{2 \cdot \gamma^2}{\gamma-1} \cdot \left(\frac{2}{\gamma+1}\right)^{\frac{\gamma+1}{\gamma-1}} \cdot \left(1 - \left(\frac{P_e}{P_c}\right)^{\frac{\gamma-1}{\gamma}}\right)} + \frac{P_e - P_a}{P_c} \cdot \frac{A_e}{A_t} = \frac{F}{P_c \cdot A_e} \quad (8.29)$$

with P_a the atmospheric pressure (Pa) and A_e the ejection plane area (m^2). This coefficient is important, as it is a function of only the nature of the gases and of the efficiency of the nozzle.

Losses

The two previous coefficients are purely theoretical and (separately) characterize the reaction inside the combustion chamber and the acceleration of the gases in the process through the nozzle of the engine. In reality, both processes are less than perfect and encounter some losses, leading to performances lower than foreseen in theory. There may be losses associated with incomplete or not fully-stabilized combustion, thermal exchanges losses, friction losses as well as those linked to non-ideal expansion in the nozzle.

Practically, each of the previous coefficients is affected by an efficiency coefficient which has to be taken into account when determining the real performance of a rocket engine. Typical values of such efficiencies are $\eta_{c^*} = 0.99$ indicating the combustion efficiency for liquid oxygen/liquid hydrogen (LOX/LH₂), and $\eta_{C_F} = 0.975$ indicating the nozzle quality.

8.6.1 Liquid Propulsion

While the origins of solid propulsion can be traced back more than 2,000 years, liquid propulsion is a rather recent development in the field of rocket propulsion. Although Konstantin Tsiolkovsky (1857–1935) published in 1903 about the application of a liquid oxygen/liquid hydrogen,

LOX/LH₂-fuelled engine in multi-stage rockets, it was not until 1926 that Robert Goddard (1882–1945) launched the first liquid fuelled rocket. This was quickly followed by rocket enthusiasts in Germany, Hermann Oberth (1894–1989) and Wernher von Braun (1912–1977), and Russia, Friedrich Zander (often transliterated Fridrikh Tsander, 1887–1933), Sergei Korolev (1907–1966), and Valentin Glushko (1908–1989). Interestingly, all of them used liquid oxygen as oxidizer but different hydrocarbons fuels: Goddard, gasoline, von Braun, alcohol, and Korolev's and Zander's group in Moscow at the Group for the Investigation of Reaction Propulsion (GIRD) a gelled gasoline and thus this engine has to be considered the first hybrid rocket engine (1933).

The first LOX/LH₂ engine ever flown was the American RL-10, an expander cycle engine with 7 tons of thrust that was used in 1963 in the Centaur second stage of an Atlas launcher. There were three competitive Moon launcher proposals in the Soviet Union, UR-700, Yangel's R-56, Chelomei's UR-700 and Korolev's N1. The N1 was finally chosen and its N1-L3 M version foresaw the RD-57, a 40 tons thrust closed cycle engine for the third and fourth stages. As of 2013, Russia holds the records for the most powerful hydrocarbon and storable engines ever build; the closed cycle RD-171 and RD-275 engines with about 835 and 175 tons of thrust. The Space Shuttle Main Engine (SSME) still leads the charts of closed cycle cryogenic engines with 218 tons thrust. Further discussion of liquid propellant rockets can be found in [Chap. 11](#).

8.6.1.1 Application Domains and Propellants

Liquid propellant rocket engines are used for all types of applications in rocketry: for booster, sustainer, and upper stage engines, for reaction control, for apogee purposes in satellite delivery, and for satellite propulsion and attitude control. Hence, the thrust level span of liquid rocket engines ranges from almost 10 MN for booster engines down to 1 N for attitude control, and the growing interest for micro-propulsion systems has led to the development of millinewton propulsion systems. The thrust of an engine is proportional to the mass flow rate going through it and to the exit velocity, which (among factors such as the propellant combination, which defines the heat released, and the weight of the exhaust gases) depends on the ambient pressure. Hence, booster engines that are supposed to provide the necessary high-thrust at liftoff are characterized by large mass flow rates at moderate specific impulses. With decreasing ambient pressure and significantly reduced thrust requirements at higher altitudes, upper stage engines usually have rather limited mass flow rates but should operate at the highest possible specific impulse because they burn propellant that has already been accelerated, and is thus of considerable value. This is the main reason that upper stage

engines usually operate with LOX/LH₂, which has the highest specific impulse of the traditional propellant combinations.

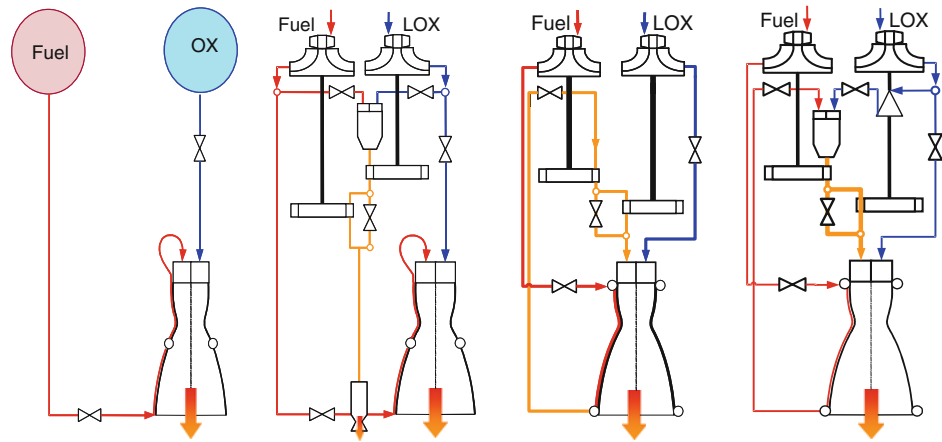
Although there is a large number of possible propellant combinations, only a few are operational. The oxidizers include liquid oxygen (LOX) and dinitrogen tetroxide (N₂O₄); typically referred to simply as nitrogen tetroxide, NTO. Kerosene fuel derivatives of hydrazine include monomethyl-hydrazine (MMH) but this is principally restricted to upper stages because it costs approximately three times that of the unsymmetrical dimethyl-hydrazine (UDMH) that is often used for boosters. Pure hydrazine is mainly used as a monopropellant for attitude control, or sometimes associated with NTO for small apogee engines. The other common fuel is liquid hydrogen. Generally, high-density propellants are used for booster applications, i.e. RD-275, RD-180, while LOX/LH₂ are either core engines, i.e. SSME, LE-7A or Vulcain 2, or upper stage engines such as the RL-10, HM7B, LE-5A or RD-0124. Typical sea-level specific impulse values of storable engines are below 300 s; LOX/hydrocarbon engines may reach 310 s, but are still far below cryogenic engines. The SSME or the Vulcain 2, which are both designed for optimum performance at higher altitudes, provide sea-level specific impulses values around 360 and 340 s, respectively.

The propellant choice can often result from a combination of system requirements, company experience and capabilities, and recurring and non-recurring costs. Hence, upper stage engines can be found with all the propellant combinations mentioned previously, i.e. the storable pressure-fed AESTUS engine, the LOX/kerosene closed cycle RD-58 M engine, and the cryogenic expander cycle RD-0124 with vacuum specific impulses of 324, 353, and 459 s, respectively. The main advantage of storable propellants is their hypergolic nature, which provides for engine ignition without any additional subsystem and offers even re-ignition capabilities. Engine re-ignition after a cruise phase is mandatory for the release of multiple payloads into different orbits.

8.6.1.2 Engine Thermodynamic Cycles

Liquid propellant rocket engines are divided into two categories depending on the type of fuel feeding technology: pressure-fed or pump-fed. The first are the simplest, the latter require additional subsystems such as gas generators or preburners, turbopumps or heat exchangers. Typical pressure-fed engines have a rather lower combustion chamber pressure which limits both the attainable specific impulse and the thrust because they use gas stored in high-pressure tanks (typically 30 MPa helium). Most of the engines used in launchers work with turbomachines to provide the required mass flow rates at the design pressure to the combustion chamber. Pump-fed engines can be

Fig. 8.9 Flow schemes of different thermodynamic engines cycle, from *left to right*: pressure-fed system, gas generator, expander, and fuel-rich staged combustion cycle



distinguished according to the method they use to generate the hot gases that drive the turbines.

Open cycle engines burn some of the propellant in a gas generator that quite often operates fuel-rich in order to limit the temperature of the combustion products to acceptable values, and they discard the turbine exhaust gases at low pressure—a procedure that lowers the overall performance of the engines since it does not use all the fuel to generate thrust. Such engines are limited in combustion chamber pressure to about 12 MPa and they therefore, tend to become rather large at high-thrust levels. Nevertheless, they have the advantage that the interface between the thrust chamber and the turbopump is simple, their components and subsystems can be developed independently, and they have a rather moderate pressure level throughout the system. A pressure-fed system is shown in Fig. 8.9 alongside three different pump-fed systems: an open cycle gas generator, an expander cycle, and a fuel-rich staged combustion cycle.

Closed cycle engines generate their driving gas in a preburner similar to open cycle engines, but the pressure level at the turbine exit is still sufficient for the preburned gases to be injected into the main combustion chamber. Thus, the entire propellant mass generates thrust. Obviously, such an engine has the potential to provide the highest possible performance and allows for a compact design. The performance of a staged combustion (closed) cycle is compared in Fig. 8.10 against a gas generator (open) cycle. While at lower pressure the performance differences cannot justify the higher complexity of a closed cycle engine, the performance advantage becomes obvious at higher pressures where the open cycle engines runs into a limit due to the increasing losses from the secondary flow that is necessary to drive the turbines. Nevertheless, the closed cycle systems require a complex development since the subsystems are highly coupled. Additionally, engine transients are much more complex due to the generally high pressures and mass flow rates, and the multiple combustion devices and components.

A third version of a pump-fed engine is the expander cycle. Here the driving gases are not produced in combustion devices, but are generated by heat addition to one of the propellants during the cooling of the combustion chamber. Similar to the open cycle gas generator, the expander cycle suffers from a limitation of the attainable combustion chamber pressure. The main advantage of this cycle is a simpler engine design, as it does not require another combustion chamber, and its ignition system. Furthermore, such an engine more readily fulfills the requirements for re-ignition; because it lacks the second ignition system, and neither the pipes nor the turbines and other components become contaminated with combustion exhaust gases during prior use.

Among the pump-fed cycles, there exist different versions; for example, there is an expander bleed cycle where only a small portion of the overall fuel mass flow rate is heated in the cooling channel of the engine and is then dumped as a film into the nozzle extension. The drastically increased pressure ratio across the turbine allows for higher combustion chamber pressures and an increased nozzle expansion ratio at reduced size and weight, almost totally compensating for the performance losses due to the partially open cycle operation.

While all flying staged-combustion engines operate oxidizer-rich (all Russian LOX/kerosene engines) or fuel-rich, a fuel-flow cycle engine in which both propellants are partially preburned would be optimum from a system point of view.

8.6.1.3 Thrust Chamber Assembly: Injection System

The thrust chamber assembly consists of the propellant manifolds, the injection head, the ignition system, the combustion chamber, and the thrust nozzle. The main purpose of the propellant manifold is to homogenize the incoming propellant and provide an even flow to the injection system.

The injection system of a rocket engine has to fulfill a rather broad range of requirements. It serves a two-way

Fig. 8.10 Comparison of staged combustion (*closed*) and gas generator (*open*) cycle engine performance

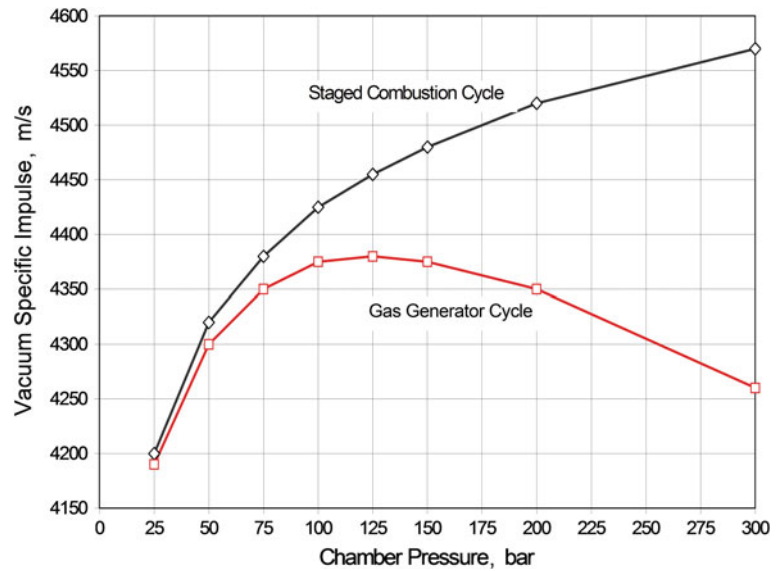
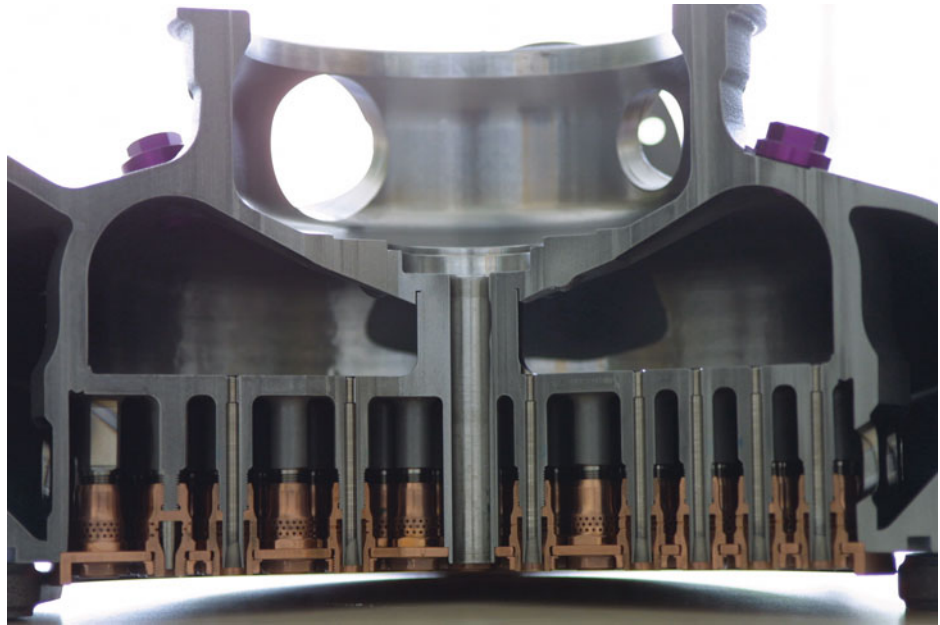


Fig. 8.11 Injector head of the Vinci engine including propellant distribution manifolds of oxygen and hydrogen and coaxial injectors. Image ASTRIUM GmbH



decoupling function, decoupling both the combustion chamber from the propellant supply lines, and the propellant manifolds from the combustion chamber in order to ensure sufficient margin from low-frequency combustion instabilities. It also has to provide a homogeneous distribution of the propellant in the combustion chamber for an optimum combustion efficiency, whilst providing all this with a minimum pressure loss for both, damping and atomization, and mixing, which amount typically to about 15–25 % of the combustion chamber pressure. Depending on the combustion device (main chamber or gas generator), the engine cycle, the propellants, and more specifically on the phase of the fluid, the injectors can be characterized as gas/gas or gas/liquid or liquid/liquid injectors.

Shear coaxial injectors, which now usually operate without swirl inducing elements are commonly used for gas/liquid applications where one of the propellants, mostly the hydrogen, is used as coolant in the combustion chamber, and which in the case of hydrogen then enters the injector as cold supercritical gas.

For open and closed cycle cryogenic engines, shear coaxial injectors are applied all over the world and can be considered as a standard injector that provides excellent atomization and mixing, and combustion efficiencies in excess of 99 %. The injector head of the European upper stage engine Vinci, developed by Snecma, I shown in Fig. 8.11, which shows the oxygen upper distribution manifold, the propellant separation plate, which includes

the liquid oxygen injectors with the metering orifices at the entrance, hydrogen distribution manifold, and the copper inserts that homogenize the hydrogen inflow into the coaxial ring.

The classical liquid/liquid injector is an impinging injector that exists in a large number of variations. They are all designed to form a liquid sheet prior to the atomization, and are characterized as like-on-like when jets of the same propellant type impinge on another, or like-on-unlike when the fuel and an oxidizer jets are onto the liquid sheet. Depending on the number of impinging jets, they are called doublet, triplets, quadruplets, and so on.

Russia has a long tradition of swirl injectors, and their engines almost always use this type of injector for liquid/liquid applications instead of the impinging injectors that are common in European and American rocket engines.

Showerhead injectors in which the oxidizer and fuel are injected as parallel jets have currently fallen out of use. Another type of injector, the American TRW pintle injector that was applied in the Apollo Moon-lander engine to achieve its deep throttling function, has recently seen growing interest due to its application in SpaceX's Merlin engines.

8.6.1.4 Thrust Chamber Assembly: Ignition System

Engine start-up and ignition are among the most severe operating conditions of a rocket engine. Malfunctions during this critical phase often result in a catastrophic failure. Hence the injector head, thrust chamber, ignition system and start-up sequence have to be designed and developed in parallel to ensure safe and stable operation.

The ignition system has to provide for the necessary energy both in time and in space, and for long enough to ignite the injected propellants. In order to fulfill this function a number of conditions have to be met simultaneously. The mixture ratio of the propellants in the chamber has to be favorable for ignition, and even more important for flame spreading and anchoring at the injectors. The energy provided by the igniter has to be high enough to ensure propellant vaporization, and to preheat them above the ignition temperature. With one or both propellants injected at around 100 K this heat transfer is critical for the desired engine start-up. Typically, an ignition delay of several milliseconds will often yield an accumulation of enough propellants which, when they react, will result in pressure peaks that can be dangerous for the combustion chamber itself, or another component of the engine cycle. Additionally, such pressure peaks can trigger combustion instabilities that can result in damage to the component and loss of the entire mission.

The majority of the ignition systems flying today are solid propellant charges that provide the necessary initiation energy to the combustion chamber. A few systems use hypergolic assisted ignition, a method where for a short

period of time a third type of liquid is injected, which reacts without further energy input with one of the propellants. Finally, for upper stage engines such as the European Vinci engine, electric igniters are used.

8.6.1.5 Thrust Chamber Assembly: Combustion Chamber Liner

The main objective of combustion chamber design is to burn completely the propellants, and to accelerate the exhaust gases to sonic velocities in the throat. Design difficulties are precise and it is difficult to make reliable predictions of optimum liner contour, combustion efficiency, hot gas side and coolant side heat transfer, and appropriate cooling, combustion chamber life, and, of course, reliable and justifiable requirements and interface conditions to other components. The combustion chamber consist of a relatively short cylindrical part and the throat area with a converging and diverging part. The throat section is where the highest thermal loads in the chamber are reached. The diverging part of the chamber extends down to expansion ratios of about 5–8, and is integrated into the combustion chamber using similar materials and cooling philosophy. The thrust nozzle extension is a separate component often fabricated using a different material as well as a different cooling cycle, for example, film cooling, dump cooling or radiation cooling. The key challenge of the integrated combustion chamber and nozzle design process is to predict the cooling system behavior and combustion performance.

The heat transfer from the hot gases to the coolant is coupled via the heat conductivity in the liner structure. It is worthwhile noting that in the typical operating conditions of rocket engines almost all of the thermo-physical properties of fluid and wall materials are not constant and are functions of the local temperature. The entire heat transfer problem can only be solved in a fully coupled manner, but a coupled solution based solely on numerical tools is not currently feasible and cannot be expected in the near future. The reasons for this are both numerous and serious, including in particular the differences in length scales of the combustion chamber, injector element, and boundary layer issues, the time scales of non-equilibrium thermodynamics, finite-rate chemistry, the presence of areas with sub-, trans-, and supersonic velocities, atomization in general and atomization under sub-, trans-, and supercritical conditions. The necessity to implement a complex thermodynamic description of processes, specifically the properties and behavior of gases, liquids and solids under cryogenic conditions, adds further complexity. Furthermore, dissociation of the exhaust gases has a direct impact on the combustion efficiency. And finally, catalytic reactions at the surface may additionally influence the local heat balance and thus influence the overall heat transfer. In the case of hydrocarbon fuels, decomposition reactions due to pyrolysis in

the cooling channels may increase the complexity of the coupled problem even further.

With all that said, it is rather obvious that semi- or fully empirical correlations in the form of Nusselt relations can be used. A typical example is the Bartz equation, which describes the hot gas side heat transfer

$$Nu = 0.062 Re^{0.8} Pr^{0.3} \quad (8.30)$$

where, $107 < Re < 108$, and $Pr \sim 0.5$. Various modifications of this basic relation have been used to make use of known or measured local geometrical quantities or thermodynamic and fluid properties that influence the local heat transfer, in order either to improve the predictive capabilities of the relation for a given set of operating conditions or to enlarge the parameter range of their application. For example

$$\alpha_g = \left[\frac{0.026}{D_t^{0.2}} \left(\frac{\mu^{0.2} c_p}{Pr^{0.6}} \right)_{ns} \left(\frac{(p_c)_{ns} g}{c^*} \right) \left(\frac{D_t}{R} \right) \right] \left(\frac{A_t}{A} \right)^{0.9} \sigma \quad (8.31)$$

with

$$\sigma = \left[\frac{1}{2} \frac{T_{wg}}{(T_c)_{ns}} \left(1 + \frac{k-1}{2} M \cdot a^2 \right) + \frac{1}{2} \right]^{-0.68} \left[1 + \frac{k-1}{2} M \cdot a^2 \right]^{-0.12} \quad (8.32)$$

and

$$Pr = \frac{4k}{9k-5} \quad (8.33)$$

and $\mu = (46.6 \cdot 10^{-10}) M^{0.5} T^{0.6}$.

For the coolant side heat transfer, design engineers generally use similar types of Nusselt correlations that vary with the operating conditions of the engines. The most sophisticated try to quantify the influence of geometry, chemistry, thermodynamic and fluid dynamic state in the form of products of non-dimensional relations with varying coefficient in the form of

$$Nu = K Re^a Pr^b \left(\frac{\rho}{\rho_w} \right)^c \left(\frac{\mu}{\mu_w} \right)^d \left(\frac{k}{k_w} \right)^e \left(\frac{\bar{c}_p}{c_p} \right)^f \left(\frac{p}{p_{cr}} \right)^g \left(1 + \frac{2D}{L} \right)^m \quad (8.34)$$

All these correlations are based on different experiments by different working groups, and therefore the results and finally the coefficients obtained depend on the experimental setup, the facilities, the operating conditions, and the measuring techniques applied, and consequently include all the known and unknown errors.

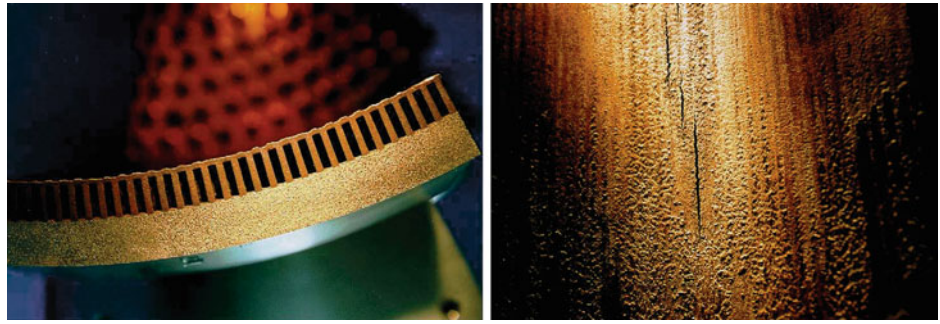
8.6.1.6 Thrust Chamber Assembly: Cooling and Thrust Chamber Life

The standard cooling method in rocket engines for launcher applications is regenerative cooling, which is sufficient for chamber pressures of up to roughly 100 bars. One propellant, typically the fuel, enters through a distribution manifold at the end of the combustion chamber and flows in the counter-flow direction upstream through cooling channels to the injector head, thus entering the combustion chamber preheated through appropriately designed injectors. Both the hot gas side and the coolant side heat transfer that defines the overall thermal loads and the cooling channel pressure loss are extremely dependent on the cooling channel design. However, at higher pressures and heat loads it is common to apply additional film cooling using cold fuel. Such a film not only acts as a coolant, which reduces the heat pick-up requirement in the cooling channels and therefore helps to reduce the pressure losses, but also protects the walls in the injector region from chemical attacks. Film cooling as the sole cooling method can only be applied in low-pressure satellite engines. Ablative cooling may be seen as a special form of film cooling, and has been applied either in the throat region of medium-pressure short burn-time booster engines, or in low-pressure satellite engines. The functional principle is based on the interaction of the following phenomena: heat input from the hot gas flow leads to a melting and vaporization of the wall material and this, together with an endothermic reaction, establishes a near-wall coolant film. Materials used are carbon, carbon/carbon (C/C) or silicon carbide (SiC) structures without infiltrated hydrocarbons.

While combustion chambers are cooled in a closed cycle mode, thrust nozzles are either cooled in an open cycle mode called 'dump cooling' or through radiation cooling. Obviously, this method relies on materials that can withstand high temperatures and is thus only applicable in satellite engines or in extension of thrust nozzles. Independent of the material applied, refractory metals such as tungsten, W, rhenium, Re, or iridium, Ir, or ceramic matrix composites, the surface of thrusters that work within the atmosphere has to be equipped with an oxygen protection layer. Upper stage engines such as the RL10-B or the new Vinci usually use a ceramic matrix composite (CMC) nozzle extension. It is worth noting that radiation cooled thrusters or nozzles require a shield system that protects sensitive engine parts or measurement equipment from high heat loads.

Thrust chambers are exposed to high thermal and mechanical stresses coming from internal and external sources. Mechanical loads resulting from high pressures in the cooling channels and the combustion chamber are combined with thermally induced loads resulting from steep temperature gradients and large temperature differences in the liner walls. During transient operation, additional

Fig. 8.12 Cut through a combustion chamber liner (*left*), liner wall with blanching and cracks (*right*). Image ASTRIUM GmbH



mechanical loads may result from flow separation induced side loads in the nozzle or through gimbaling of the engine. These extreme loads can lead to two different failure modes: fatigue and rupture. Repeated cyclic loads, which are usually characterized according to the frequency of their occurrence as either low cycle fatigue (LCF) or high cycle fatigue (HCF), contribute to the fatigue that is additionally aggravated by high temperature creep of the material under plastic strain over an extended period of time (creep fatigue). There is another damage effect described as ‘blanching’, which is caused by a chemical attack of oxygen radicals of the liner surface in the vicinity of the injector face plate. Some of the loads described previously become more severe during engine start-up and shutdown when extreme temperature gradients are reached and the thermal stresses are coupled with transient mechanical loads. The classical fatigue dominated failure mode is a rupture of the cooling channel called the ‘dog-house’ effect. The left part of Fig. 8.12 shows a cut through the combustion chamber liner with the beginning of the dog-house effect visible through a roughening of the hot gas side wall. In the right part of Fig. 8.12 the inner liner with the typical combination of blanching and cracks after about 20 cycles and more than 15,000 s of operation is shown. The consequence of these combined loads is to make it almost impossible to build, with current materials and manufacturing techniques, a high-pressure and high-performance reusable cryogenic rocket engine.

8.6.1.7 Thrust Chamber Assembly: Oscillations and Combustion Instabilities

Extremely high energy densities in combination with very low internal losses make a rocket engine a nearly undamped system with high amplification, and thus very sensitive to oscillations, the sources of which can be diverse. For example, interactions of cavities in the propellant supply lines and the rocket engines, or interaction between the propellant distribution manifold and the injectors with the combustion chamber and combustion noise itself.

Low frequency induced oscillations, which result from resonances between of stage structures, propellants and the engine are called ‘pogo oscillations’ and, depending on the size of the vehicle may vary around 100 Hz. Design

engineers know how to avoid them by installing anti-pogo devices into the fluid lines between the tanks and the engines to act as dampers. Furthermore, there is another type of low frequency oscillation between the propellant lines and the engine, a phenomenon called ‘chugging’ whose typical frequencies may vary between 100 and 700 Hz. Quite often, pressure-fed engines are more vulnerable to this problem as they are designed for very low pressure losses and thus have only marginal internal damping. A major requirement for injectors of liquid propellant rocket engines is to provide for sufficient losses in order to increase the damping in the system.

Oscillations inside the combustion chamber can result from a coupling of the heat release with the eigenfrequency of the combustion volume. Resonances may occur within these oscillations in preferred frequency bands, and give rise to combustion instabilities, a process which is characterized by very high pressure amplitudes that can destroy hardware within only a few acoustic cycles. These high frequency oscillations continuously disrupt the boundary layer and bring hot combustion gases to the walls, which in combination with the high pressure peaks, produces local heat loads that can significantly exceed the material’s limit.

Although combustion instabilities have been the subject of extensive research within almost every development program of large rocket engines, the lack of an exact mathematical formulation of the phenomenon persists. Nevertheless, design engineers have found ways to tackle this problem and either bypass it by implementing passive means such as acoustic cavities in the combustion chamber liner placed in the vicinity of the face plate to modify the eigenfrequency of the chamber, or by placing baffles inside the combustion chamber to subdivide the volume into smaller ones with different and higher characteristic frequencies. Furthermore, it is known that engines that burn hydrocarbon propellants are more sensitive to combustion instabilities than LOX/LH₂ engines, and that storable propellant engines, which quite often operate with impinging injectors, show an even higher sensitivity. Additionally, the likelihood of instabilities increases with the diameter of the combustion chamber and is prevalent during the transient start-up and shutdown processes of the engines. Russian engineers of high

thrust engines often implement a design with multiple small combustion chambers supplied by a single turbomachinery.

It is noteworthy that any newly developed rocket engine has to undergo a well-defined experimental program that includes bomb tests in which small solid charges are placed inside the chamber and ignited during engine operation to trigger a steep pressure increase and then determine the damping characteristic of the engine.

8.6.1.8 Thrust Chamber Assembly: Nozzle Technologies

Thrust nozzles have to be designed such that the hot combustion gases are accelerated to the maximum exhaust velocity. The exhaust flow generated by a bell-shaped nozzle with a parabolic contour is almost ideal with only a minor portion of non-parallelism. Typically, Russian nozzles have an ideal contour, which follows a classical wind tunnel design and which is adapted to the ambient pressure, while American, European and Japanese engines follow a thrust-optimized contour, which has a somewhat steeper opening near to the throat and yields a shorter nozzle at similar expansion ratio. It is of note that thrust-optimized nozzles are in general more sensitive towards side loads during operational transients. Hence, they require a somewhat more robust design that may, to some extent, counteract the weight advantage.

Classical booster engines, which typically operate for less than 3 min are less sensitive to flow separation induced side loads as their expansion ratio usually does not exceed 20–30. Core engines such as the Japanese LE-7 or the European Vulcain 2, which are ignited on the ground and operate for almost 10 min, have to be designed such that at sea level the nozzle operates in an over-expanded mode, meaning that the exhaust pressure is substantially lower than the ambient pressure, typically between 20 and 30 %. During transient start-up, flow separation may cause severe side loads and therefore the nozzle design is a compromise between safe start-up and high performance at high altitude.

There are at least four nozzle concepts available: the plug nozzle, the expansion-deflection nozzle, the dual bell nozzle, and the extendible nozzle, which one way or the other aim to adapt the flow to the ambient pressure. Two well-known upper stage engines favor the concept of the expandable nozzle, the American RL-10B and the European Vinci, which both feature a ceramic matrix composite (CRC) nozzle extension. This is however, only for the purpose of reducing the height of the launcher, and the nozzle extension is deployed after the separation of the first stage, and prior to engine ignition.

8.6.1.9 Thrust Chamber Assembly: Manufacturing Issues

The design of a combustion chamber depends almost entirely on the applied cooling technologies described

previously. While low pressure engines might as well be manufactured from film cooled and thermal barrier coated steel, and short burning time engines such as the Viking engine might contain ablative inserts, the state-of-the-art of high pressure combustion chambers in the Western world is a high heat conductivity inner liner that is combined with a high strength outer shell. The inner copper liner includes the cooling channels that are milled into the copper and then filled with wax and closed either by electro-deposition or by being brazed to the outer nickel shell. The latter technology is widely used in Russia and has been applied to the American RS-68 engine. The earlier tube design of the inner liner is limited to rather low combustion chamber pressures and is currently used only in low pressure upper stage engines such as the American RL-10.

8.6.1.10 Thrust Chamber Assembly: Gas Generators and Pre-burners

The main purpose of a gas generator, or in the case of a staged combustion cycle engine a preburner, is to provide the driving gases for the turbomachinery. The turbine power requirement depends on the mass flow rate and thermodynamic properties (c_p , κ , T_1) of the driver gases, on the turbine efficiency, and the available pressure ratio (p_2/p_1) across the turbine. Thus

$$P = \eta \dot{m} c_p T_1 \left(1 - \left(\frac{p_2}{p_1} \right)^{\frac{\kappa-1}{\kappa}} \right) \quad (8.35)$$

All operational open cycle engines have gas generators running fuel-rich in order to limit the turbine entry temperature independently of the propellant combination; e.g. the cryogenic Vulcain, the LOX/kerosene RD-180, or the storable Viking engines. In contrast to closed cycle engines, in open cycle engines the turbine exit pressure p_2 and the combustion chamber pressure are not directly coupled. Hence, open cycle gas generator pressure and the main chamber pressure are typically in the same range, whereas in closed cycle engines the preburner pressure may exceed the chamber pressure by a factor of two. Classical closed cycle engines such as the cryogenic SSME or the LOX/kerosene RD-180 have chamber pressures of about 22.5 and 26.7 MPa, respectively and preburner pressures of 360 and 540 bars respectively. Due to these already extreme mechanical and dynamical loads, the turbine entry temperatures typically do not exceed 900 K, and in order to avoid local thermal overloading of the turbine blades a homogenous temperature distribution is extremely important. While all preburners of cryogenic closed cycle engines operate in fuel-rich mode to limit the turbine entry temperatures, preburners of LOX/kerosene engines operate oxidizer-rich to avoid soot formation and

deposition either on the turbine blades or in the injection system.

The injectors used in fuel-rich gas generators are often similar to those applied in the main chamber because the propellant temperatures are only slightly different. In oxidizer-rich preburners, which are much more sensitive to combustion instabilities, a two-stage combustion zone is developed. In the first part of the chamber, a limited amount of oxygen is injected in order to keep the temperatures above 2,000 K, and in the second part, the remainder of the oxygen is injected to dilute the hot gases sufficiently to reach the desired turbine entry temperature. This secondary injection also helps to achieve a homogenous temperature profile. Problems of material compatibility of gas generator, fluid supply, and the turbine against hot oxidizer-rich gases can be overcome either by passivation of the surfaces or by an appropriate coating.

8.6.1.11 Thrust Chamber Assembly: Test Facilities

Within the development phase of an engine, all propulsion system components are usually tested in parallel and only later qualified together as subsystems or engines using specific test facilities. Such a development and testing methodology requires various large-scale facilities that are a major cost driver.

System tests facilitate characterization according to the engine type tested in sea-level and altitude simulation facilities. Booster and main stage engines are tested at sea-level conditions, and upper stage, apogee, and satellite engines at high-altitude conditions. A characteristic feature of altitude simulation benches are the systems and installations required in order to establish and maintain vacuum-like conditions during engine firing.

There is a general rule: fly as you test and test as you fly, and no component, subsystem or engine will ever fly before it has been demonstrated to fulfill the desired requirements and safety margins. However, it is not always possible to totally realize this principle. For example, the ambient pressure during ascent decreases continuously, thereby changing the thrust as well as the pressure difference across the thrust nozzle. It is extremely costly to modify a test bench for large rocket engines to simulate the ascent, so such tests are often omitted. Additionally, installations typical for the launch site such as the tower, the water-cooling, or the operation of additional engines (e.g. large solid boosters) are nearly impossible to realize on a test stand.

8.6.1.12 Turbopumps and Turbines

The thrust of an engine is directly proportional to the ejected mass flow, which in turn is directly proportional to the combustion chamber pressure

$$\dot{m} = \frac{P_c \cdot A_t}{c^*} = C_{\dot{m}} \cdot P_c \cdot A_t \quad (8.36)$$

with \dot{m} the mass flow (kg/s), A_t the throat area (m²), c^* the characteristic velocity (m/s), $C_{\dot{m}}$ the mass flow coefficient (1/s), and P_c the combustion pressure (Pa).

Various combinations of engine cycle can be considered, as seen previously in Fig. 8.9. The pressure fed engines are the simplest, but are limited in thrust by the need for high pressures in the tanks, leading to very heavy assemblies. Practically, it is difficult to have a thrust higher than 50 kN with a pressure fed (or blow-down) cycle, which corresponds in general to a combustion chamber pressure of some 10–20 bars. To reach higher thrust values, higher chamber pressures are necessary, imposing the need for turbopumps.

The function of a turbopump is to raise the pressure of a propellant from the low pressure of the storage tank to the high pressure necessary for its injection into the combustion chamber. Turbopumps are generally rotating machines, although some manufacturers have considered piston machines similar to an automobile engine (for instance Ukrainian proposals, and XCor).

The power developed by a turbopump P_P (W) can be very high, often reaching several MW, depending on the pressure increase ΔP (in Pa), the mass flow q (kg/s), the volumic mass of the propellant ρ (kg/m³), and the efficiency of the machine η_P . Typical efficiencies are in the range 0.6–0.8 depending on the design of the pump.

The power of a turbopump

$$P_P = \dot{m} \cdot \frac{\Delta P}{\rho \cdot \eta_P} P \quad (8.37)$$

is generally provided by a gas turbine, the function of which is to expand high enthalpy gases in a rotating machine to provide mechanical power to the axis of the turbine. The power developed by a gas turbine P_T (W) is a function of the gas mass flow \dot{m} (kg/s), the thermal capacity of gas C_P (J/kg.K), the isentropic temperature T_{Isos} (K), and the efficiency of the turbine η_T . Typical efficiencies are in the range 0.45–0.7 depending on the design of the turbine.

The power of a turbine driving a turbopump

$$P_T = \dot{m} \cdot C_P \cdot \Delta T_{Isos} \cdot \eta_T \quad (8.38)$$

has to be transferred from the axis of the turbine to the axis of the pump. Numerous schemes are conceivable, depending on the number of turbines (one per pump, or one driving two pumps) and the respective rotation velocities of the various machines, these being mainly dependent on the density of the propellants.

A major difficulty to solve with turbopumps is the risk of cavitation. In order to function correctly, the inlet pressure of a pump must be high enough to avoid the creation of gaseous

bubbles whose presence can damage the blades of the pump. Define the net positive suction pressure (NPSP) of a pump (Pa) as the difference between the total inlet pressure P_t —which is itself the sum of the static pressure P_S and the dynamic pressure P_d —and the vapour pressure of the fluid P_v , viz

$$NPSP = P_t - P_v = P_S + P_d - P_v = P_S + \frac{1}{2} \cdot \rho \cdot V^2 - P_v. \quad (8.39)$$

This NPSP has to be compared to the pressure effectively available for the propellant tank, the sum of the static pressure in the tank, the hydrostatic pressure, and the pressure losses in the feed lines. When this pressure leads to a negative NPSP, or in reality to values too low to operate the pump safely, then cavitation may occur and potentially destroy the machine. One way to avoid this phenomenon is to include on the pump axis a suction stage called a boost-pump to help with the initial pressure increase. This leads to very complex and costly elements, but is a necessary evil for high performance liquid rocket propulsion.

8.6.2 Solid Propulsion

The use of black powder to propel a small incendiary rocket or firework was discovered by Chinese alchemists at the end of the first millennium. The first description of a weapon date to around 1045 ('blazing arrow'); this knowledge migrated from China to India, then to Arab countries and then to Europe where it was used many times. For example, by the Arabs in 1095 in Antioch, an ancient city on the eastern side of the Orontes River, while Joan of Arc (1412–1431) defended Orleans in 1428 using rockets. In 1660, Blaise Pascal (1623–1662) explained and formulated the principle of rocket propulsion.

The 19th century is considered the first golden age for solid rockets: in England, under the leadership of Sir William Congreve, 2nd Baronet (1772–1828), military rockets were improved using the technologies developed for fireworks and were widely used in battle. During this period, progress in the chemistry of energetic materials also brought the invention of the so-called 'smokeless powder', based on nitrocellulose, by Paul Vieille (1854–1934), while in the same years Alfred Nobel (1833–1896) obtained the first 'double base' propellant based on a mixture of nitroglycerine and nitrocellulose. Further progress has been made since then on double base propellant formulations, currently used for instance in tactical missiles. Before the Second World War, solid rocket motors were based on extruded double base propellants and a metallic case, but the caliber was limited and the grain shapes were simple. During the Second World War, a new type of solid

propellant, called 'composite' propellant was first used for JATO (Jet Assisted Take-Off) rockets, and later improved at the Jet Propulsion Laboratory through the introduction of a polyurethane binder. Composite propellants have rapidly become widespread, and today represent the baseline propellant of many solid rocket motors used for launcher applications.

Modern civilian solid motors consist of a filament wound case containing an aluminized hydroxy terminated polybutadiene (HTPB) composite propellant grain and a movable nozzle with a flex seal. Solid propulsion is a cost efficient, mature technology for small launchers and strap-on or add-on boosters and is further discussed in [Chap. 11](#).

8.6.2.1 Applications

For a launcher, the main applications are

- Large solid rocket motors used in stages that provide most, if not all, of the thrust for liftoff (e.g. Space Shuttle, Ariane 5, Titan IV).
- Strap-on boosters assisting the launch vehicle liftoff to increase its performance (e.g. Ariane 3, Ariane 4, Delta 2, Delta 4, Long March).
- Stages of small launch vehicles (e.g. Pegasus, Vega).
- Small motors for stage separation.

In addition, solid rocket motors are employed in a various other applications such as tactical missiles, strategic ICBMs, and sounding rockets.

8.6.2.2 General Description of a Solid Rocket Motor

A solid rocket motor is basically a high pressure tank containing the mass of solid propellant, called the propellant grain, suitably shaped to produce the desired pressure (thrust) time history, and in such a manner as to leave an internal volume, the combustion chamber, to accommodate the combustion products. The internal wall of the pressure tank, known as the motor case, is protected from the high convective and radiative heat fluxes by an internal thermal insulation. The combustion occurs at the surface of the propellant grain and proceeds in a direction perpendicular to the surface. Some parts of the propellant grain may be covered by inhibitors, in order to obtain the required pressure–time curve. The gases produced by the combustion process are exhausted through a nozzle at one end of the motor case. The initial ignition of the propellant surface is obtained by way of a dedicated device, the igniter. A rupture disk, designed to burst at an assigned pressure, is usually placed at the aft end of the motor, both to protect the propellant during transport and assembly operations, and to facilitate the initial pressure build-up in the motor during the ignition transient. Additional structural elements, such

as skirts, are often included to transmit the thrust and for attachment purposes in multi-stage launch vehicles. A typical solid rocket motor is illustrated in Fig. 11.2.

8.6.2.3 Burning Rate and Internal Ballistics

As mentioned previously, the burning of the solid material occurs at the propellant surface and in a direction perpendicular to the surface itself. The rate of recession of the propellant surface is called the burning rate. An empirical relation between the burning rate r and the chamber pressure P_c is

$$r = a(P_c)^n \quad (8.40)$$

where a is a constant that takes into account the initial grain temperature, and n is a constant known as the burning rate pressure exponent. Values of r in solid motors for launcher applications are usually of the order of 1 cm/s. The burning rate is very sensitive to the exponent n . The relation $n < 1$ must hold in order to ensure a stable motor operation. On the other hand, if the value of n approaches zero then the burning process may become unstable, possibly leading to the extinguishment of the combustion. Typical values of n for production propellant are between 0.2 and 0.8.

An important effect is the so-called erosive burning, an increase in the burning rate mainly due to a high gas flow velocity (including metal or metal oxide solid particles). The increased heat transfer from the combustion products to the propellant grain is the dominant effect compared to the actual erosion of the solid propellant. Erosive burning is most likely to occur at the beginning of the burn time, when the internal passage is narrower and the gas velocities are higher, and also towards the aft end of the motor because the gas flow accelerates from the motor-head end towards the nozzle. To compensate for this effect and achieve a uniform combustion in the axial direction, the port area, the cross-section area of the flow channel in a motor, is increased by tapering the propellant in the flow direction. The main parameter affecting erosive burning is the ratio between the port area A_p and the nozzle throat area A_t . Typical values of the initial A_p/A_t ratio are comprised two and five. Additional enhancement of the burning rate is observed in vehicles with a high longitudinal acceleration or a high rate of spin around the longitudinal axis. The effect is more pronounced if the burning surface is at a high angle to the acceleration vector.

The rate of production of the gaseous combustion products equals the rate of consumption of solid propellant, and is therefore given by

$$\dot{m}_g = \rho_p A_b r \quad (8.41)$$

where \dot{m}_g is the gas generation rate, ρ_p is the propellant density, and A_b is the area of the burning surface. Under

steady state operation, the gas generation rate equals the mass flow rate through the nozzle, expressed as

$$\dot{m}_n = C_{\dot{m}} P_c A_t \quad (8.42)$$

with $C_{\dot{m}}$ being the mass flow factor (inverse of c^* used in liquid propulsion) at the nozzle throat area, A_t , and P_c being the chamber pressure. Note that Eq. 8.42 is the same as Eq. 8.36. Using the relationship for the burning rate, the chamber pressure can be determined as

$$P_c = \left(\frac{A_b a \rho_p}{A_t C_{\dot{m}}} \right)^{1/1-n} \quad (8.43)$$

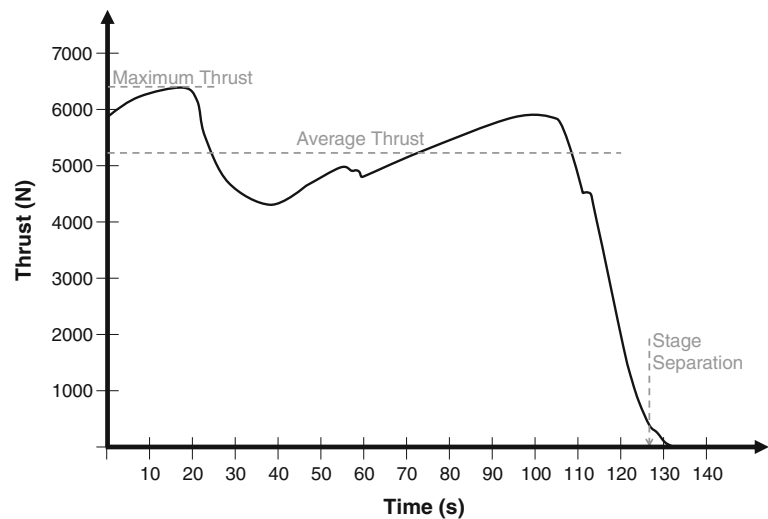
The pressure inside the motor is proportional to the burning area A_b . An increase (decrease) of burning area with time will lead to an increase (decrease) in the motor chamber pressure (and thrust). As observed, the higher the value of n , the greater is the sensitivity of the chamber pressure to small variations in burning surface. The variation of A_b with time depends on both the burning rate and the initial geometry of the propellant grain. This initial shape determines the pressure (thrust) history. That is a progressive, neutral, or regressive burning is attained depending on whether the pressure (thrust) increases, remains constant, or decreases with time; see also Chap. 11 and specifically Fig. 11.3. In practical applications, the pressure–time curve is shaped to take into account the performance requirements and constraints of the launcher system, such as reduced thrust during the transonic phase, maximum dynamic pressure, maximum acceleration, controllability during thrust tail-off, and so forth. A typical curve for the Ariane 5 booster is shown in Fig. 8.13.

8.6.2.4 Motor Case and Thermal Protection

The motor case represents the structural element of a solid rocket motor. Its design is mainly driven by the requirement to support the motor internal pressure (the MEOP: maximum expected operating pressure), but it must also satisfy other requirements of the launcher system, such as thrust transmission, general vehicle loads, thrust vector control, motor handling, and so forth. The bare motor case design and manufacturing technologies have significantly progressed in recent years. Two main approaches are typically applied

- Metallic motor case: currently adopted for large segmented solid rocket motors. Each segment may be formed by several cylinders made of steel joined together either by using clevis-tang connections or by welding. The cylinders are manufactured at the desired diameter and length using flow-forming techniques, are then usually given a heat treatment (such as quench and temper), and finally the ends are machined to obtain the inter-

Fig. 8.13 Typical Ariane 5 solid propellant motor (MPS) booster thrust-time curve



segment joints. The forward and aft domes are forged, and include suitable openings for connecting the igniter and the nozzle, respectively.

- Composite motor case: used for monolithic solid rocket motors. Motors as large as 3 m in diameter and as long as 10 m can be obtained. The composite case is obtained by filament winding (mainly carbon) impregnated by an epoxy resin, serving as a plastic matrix. The filament winding deposits filament bands around a mandrel having the final motor case internal diameter. The fibers are oriented along the direction of the principal stresses, using two or more winding angles.

The internal thermal protection has the function of protecting the bare motor case from the high convective and radiative heat fluxes to which it would be exposed after the full propellant consumption (at a given axial location), in order to maintain the temperature of the metallic or composite structure at an acceptable level. The thermal protection is usually made of a synthetic rubber filled with silica, Kevlar, glass, or microspheres. Under the effect of the high thermal fluxes, a pyrolysis of the layer of the thermal protection closer to the hot gases results in ablation of the pyrolyzed layer by the mechanical erosion of the gas flow. The design of the thermal protection (i.e. the variation of its thickness with the axial position) depends on the thermo-physical properties of the chosen material, the temperature of the gases inside the motor chamber, and the duration of exposure.

For metallic cases, the internal thermal protection is put in place through dedicated winding machines on the inner surface of the bare motor case, suitably prepared by sanding, degreasing, etc., and then subjected to a heat treatment in an autoclave to achieve vulcanization of the rubber in order to bond it to the motor case. In segmented motors, the front of each propellant segment may be thermally

insulated, and hence not participate in the combustion, so as to achieve the combustion surface associated with the desired thrust law. For composite cases, the thermal protection is wound on the mandrel, then treated in the autoclave either separately or at the same time as the curing of the composite material.

8.6.2.5 Propellant Grain

A solid propellant includes all the necessary ingredients for combustion. No external oxygen is needed. The propellant provides through its combustion the gas flow rate required for propulsion. The main properties that characterize a solid propellant are its performance and internal ballistic properties (specific impulse, density, burning rate, burning rate exponent, burning rate sensitivity to temperature), the mechanical properties required in order to sustain the pressure and shear loads as well as temperature changes, its storage capability (also referred to as shelf life), hazard properties, handling and transport characteristics, production properties, and material costs.

Solid propellants are classically divided into three main categories. Double base propellants, in which the fuel and oxidizer are contained in the same molecule, are typically based on nitrocellulose and nitroglycerine. Composite propellants are composed of separate elements: a solid fuel, a solid oxidizer, and a polymeric matrix serving as a binder. Composite modified double base propellants are double base propellants enhanced by such ingredients as ammonium perchlorate, aluminum, Research Department Explosive (RDX, also called cyclonite, hexogen, or T4; chemical name cyclotrimethylenetrinitramine or 1,3,5-trinitroperhydro-1,3,5-triazine), or HMX (origin of compound name unclear but also known as cyclotetramethylene-tetranitramine, tetrahexamine tetranitramine, or octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine).

Operational solid rocket motors for launcher applications use composite propellants with ammonium perchlorate as oxidizer and aluminum as fuel because only this type of propellant is able to satisfy modern hazard regulation required (i.e. class 1.3).

Within a composite propellant, the oxidizer contains the oxygen necessary for the combustion of the fuel. Desired properties are: high density, high oxygen balance, a heat of formation as negative as possible, good thermal stability, and low impact sensitivity. Different sizes of oxidizer particle are often used in a single propellant mix, usually with both fine and coarse particles (e.g. 10 and 200 μm), so as to include in the propellant a high amount of oxidizer. The particle size distribution, and indeed the particle shape have important effects on the internal ballistic performance, and sometimes may affect such phenomena as pressure oscillations. The most conventional oxidizer is ammonium perchlorate but ammonium nitrate is also used in some applications and more advanced oxidizers include RDX, HMX, CL-20 (hexanitrohexaazaisowurtzitane, or HNIW), ammonium dinitramide (AND) and hydrazinium nitroformate (HNF). The fuel is usually represented by aluminum particles (20–50 μm) in order to provide high density and specific impulse. Other possible fuels such as boron, beryllium, or lithium have limited applications, mainly due to their poor burning characteristics, toxicity, and difficulties in manufacturing. The oxidizer and fuel particles are embedded in the binder, which ensures the cohesion of the cured propellant. The binder participates in the combustion as a fuel. The main criteria for the choice of a binder are good mechanical properties, good compatibility with other ingredients, satisfactory casting characteristics (low viscosity and slow cross-linking), rapid curing, and good aging characteristics. Binders are typically based on polybutadienes (PB) and the three main types are polybutadiene-acrylic acid acrylonitrile (PBAN), carboxy-terminated polybutadiene (CBAN) and hydroxyl-terminated polybutadiene (HTPB). Current civilian solid rocket motors use HTPB. Advanced, energetic binders are PolyNimmo, PolyGlyn, GAP, and BAMO.

Several additives are usually included in a propellant formulation. The main additives include catalysts for modifying (increasing or decreasing) the burning rate, chemical ingredients to modify the burning rate exponent, cross-linking agents, plasticizers for improved mechanical properties, stabilizers, and so forth.

The propellant grain is the ensemble of the solid propellant itself. The initial shape of the propellant grain is a very important aspect of designing the pressure (thrust) history of the motor; see Fig. 11.3. Many different shapes have been conceived, but the most frequently used for launch vehicles are

- a. Cylindrical grain, having a constant (usually circular) cross-section along the motor axis. It provides progressive burning.
- b. Conical grain, or sometimes a conocyl (i.e. the combination of a conical and a cylindrical part). It makes the burning closer to neutral.
- c. Slotted tube (fins), having a star-shaped cross-section. It provides a regressive burning, and is often used to make available a large burning surface at motor ignition and during early operation.
- d. Finocyl configuration, the grain having a star-shape in one part (either at the motor's head-end or aft-end) and being cylindrical in the other part. It combines the advantages of both the pure cylindrical grain and the star-shaped grain.
- e. End burning grain, which burns in the axial direction only on one side surface of the grain. It has a neutral burning, but is no longer in use for practical applications in launch vehicles.

The propellant grain has to withstand several loads during its life: pressure loads (in particular during the ignition transient), mechanical loads (both gravity and accelerations), thermal loads during cool-down after curing (propellant shrinkage), thermal cycling during handling, transportation and storage, and dynamic loads during motor transportation and functioning (vibrations, shocks and acoustic noise). The propellant is a visco-elastic material whose mechanical behavior is time dependent, having the capability to accumulate damage from different load sources or repeated load conditions. The mechanical properties of the propellant grain (Young's modulus, stress, and strain) are usually plotted versus time for different values of temperature and strain rate. Laboratory analysis of the propellant strength is usually done using uniaxial loading tests, and in some cases with biaxial loading tests. The structural analysis of the grain (including the liner, thermal protection, and motor case) is performed using finite element method calculations, with the goal of verifying that acceptable margins of safety are attained. As the propellant is case bonded, special attention has to be given to areas such as the bond line or the grain ends, where high stress concentrations may occur. Floaters are applied at these propellant-case interfaces in order to compensate for stresses induced during propellant shrinkage (primarily different thermal expansion coefficients of the case material and the propellant) and by the pressure loads during motor ignition.

Propellant grain manufacturing involves complex processes that are performed under strict safety measures in order to prevent explosion or fire. The manufacturing process depends on the type of propellant and on the motor size because there is no single standard process. Composite propellant preparation is usually based on a batch process,

although a continuous-flow process also exists. Batches with the polymer and the fuel particles, together with any burning rate catalyst and plasticizers are prepared, and as it undergoes a mixing process to make it homogeneous the oxidizer is introduced. Several batches can be prepared, in which the mixture (pre-mix) remains in liquid state. Then the curing agents are added and a final mixing is performed. The propellant is poured into the protected motor case and placed in the casting pit. A mandrel is introduced inside the motor in order to achieve the final motor internal perforation, suitably shaped (cylindrical, star, etc.). The propellant hardens during the curing phase over a period of several days, performed typically at around 50 °C. Once curing is completed, the mandrel is carefully removed, leaving the propellant bonded to the internal thermal protection through a liner. The liner is made of a material based on the same polymer as that of the propellant to ensure good compatibility, and is applied through a process of spray deposition in order to control its thickness.

It is essential to avoid cracks, voids and any type of flaw in the propellant that would cause off-nominal propellant burning and motor functioning, in some cases possibly leading to catastrophic failure. Several techniques are employed to this end, including the application of vacuum and temperature control during mixing and casting, vibration (and sometimes spinning) of the motor case during casting, and slight motor pressurization after casting in order to remove air bubbles. Finally, non-destructive inspection techniques, typically X-rays and ultra-sound, are employed to ensure propellant and motor integrity. After these controls, the loaded motor case can be completed with the integration of the nozzle and the igniter.

8.6.2.6 Igniter

The igniter's function is to ensure the proper ignition of the propellant. Several types of igniter exist, but the most widely used is the so-called pyrotechnic igniter assembly. It usually consists of a two or three igniters operating in a chain to generate at each stage an increasing mass flow of hot gases until reaching a mass flow level sufficient to reach the ignition conditions in the propellant grain. As an example, the first stage of the chain is always a pyrotechnic igniter consisting of pellets (usually containing boron, potassium, or magnesium), mounted inside an intermediate igniter, which in turn is placed inside the main igniter. Each igniter of the chain may contain several nozzles, canted at a predetermined angle with respect to the motor axis in order to ease ignition. The pyrotechnic igniter is started by an initiator, the IFOC (Inflamateur à Fonctionnement par Onde de Choc; Detonation to Deflagration Initiator), also known as a squib or primer, delivering the initial energy upon receipt of an electric signal. The whole igniter assembly is placed on the motor-

head end dome through a bolted flange. A safe and arm device is frequently used to prevent inadvertent motor ignition.

Motor ignition is a very rapid process, typically completed in less than 0.2 s. The ignition transient is conventionally subdivided into three phases: the ignition time lag, that is the time from the initial IFOC signal to the first ignition of a point in the motor propellant grain; the flame spreading interval, that is the time from first ignition of the propellant grain to the full ignition of its burning surface; and the chamber filling interval, that is the time for filling the combustion chamber with hot gases, during which the chamber pressure increases until attaining an equilibrium value. The ignition transient is a critical phase in a motor's operation: a pressure peak above the equilibrium chamber pressure may occur in some rocket motor designs during the chamber filling interval. Also, in launch vehicles with two or more boosters, the ignition of the boosters must occur simultaneously, with only a very small difference in ignition transient duration being allowed in order to avoid thrust imbalances. The problem of thrust imbalance during the entire motor operation is an important issue to be taken into account during the motor design phase, with the goal being to achieve good motor reproducibility and ensure the correct flight of the launcher.

8.6.2.7 Nozzle

The task of the nozzle is to evacuate the combustion gases and to generate the thrust. It is attached to the rear dome of the motor case. The classical design consists of a submerged nozzle with a conical divergent section, and it can be either fixed or movable. In this latter case, it can be orientated through a flexible bearing to allow the thrust vector to be directed (deflection angles are usually up to 6° or 8°). The nozzle is designed to support the loads resulting from the high temperature and pressure of the combustion gases and by the nozzle gimbaling.

The nozzle consists of a divergent assembly, with a metal part (in aluminum or steel) to support the mechanical loads, and a thermal protection made of either carbon phenolic or silica phenolic to provide the thermal barrier. The thermal protection is usually glued to the metallic casing. The divergent structure may be made of filament reinforced plastic in some applications. The nozzle throat (usually a throat insert) is made either of C/C or graphite. The nozzle nose forming the convergent section and nose cap is also made of carbon phenolic. The flexible bearing (or flex seal) is a sandwich of shims (metallic or glass epoxy composite) with rubber pads, having a spherical shape in order to allow rotation by introducing shear displacement in the pads. A low modulus rubber is usually employed for the pads, in order to decrease the torque and therefore the power required by the

thrust vector control (TVC) system for nozzle actuation. The flex seal is protected by a membrane.

An important effect during motor operation is the erosion of the nozzle throat due to a combination of oxidation of the carbon at high temperatures, and the rapid flow of aluminized gases. This erosion increases the throat diameter and thus alters the motor performance. It is therefore very important to use low erosion materials for the nozzle insert. The correct estimation of the nozzle throat erosion is essential for the good design and operation of the nozzle. The best materials are Pyro-Graphite or high density C/C. A movable nozzle includes the necessary connection (usually made of steel) with the thrust vector control system for swiveling the nozzle.

8.6.2.8 Pressure Oscillations

Pressure oscillations inside the combustion chamber are observed in some solid rocket motors, prevalently in large boosters such as the Space Shuttle's SRM, the Titan SRMU, and the Ariane 5 SRM. Although they rarely lead to catastrophic motor failure, these pressure oscillations translate into thrust oscillations (through the response of the motor structures) that may affect motor performance and produce high vibration loads on the vehicle structures and on the payload.

There are two main types of pressure oscillations phenomena

- Oscillations that sustain themselves through energy coupling with the unsteady combustion process. Given that the burning rate is influenced by the combustion pressure and flow velocity, oscillations in pressure (flow velocity) can interact with the propellant combustion energy release in a manner that produces self-sustained pressure oscillations.
- Oscillations sustained by unsteady hydrodynamic phenomena. In particular, the formation and shedding of vortices within the combustion chamber is responsible for driving pressure oscillations, when the shedding frequency is close to one of the main acoustic frequencies of the combustion chamber. The vortex shedding may be generated by inhibitor rings that protrude into the main chamber flow during propellant combustion, or by a rearward facing step in the propellant grain, or even by the unstable transition of the velocity boundary layer along the burning propellant surface.

In both types of pressure oscillation, the acoustical damping is mainly due to the nozzle, viscous flow effects, and particulate damping.

Despite the significant progress made in understanding the physical phenomena, and in the associated modeling, combustion instabilities and pressure oscillations in solid rocket motors continue to be an important issue in the design and operation of a motor. A complete knowledge of the physical processes is still lacking; as a matter of fact, it

has been sometimes observed that even a small variation in one of the motor parameters can have a large, unexpected impact on pressure oscillations. Work is continuing to improve understanding of the main parameters affecting pressure oscillations, to enhance validated simulation tools and provide reliable estimations not only of the oscillation frequencies but also of the oscillation amplitudes. This will make available flexible test motors at an intermediate scale that are, sufficiently representative of the involved phenomena to preclude the need to construct and operate full-scale solid motors.

8.7 Launcher Avionics and Software

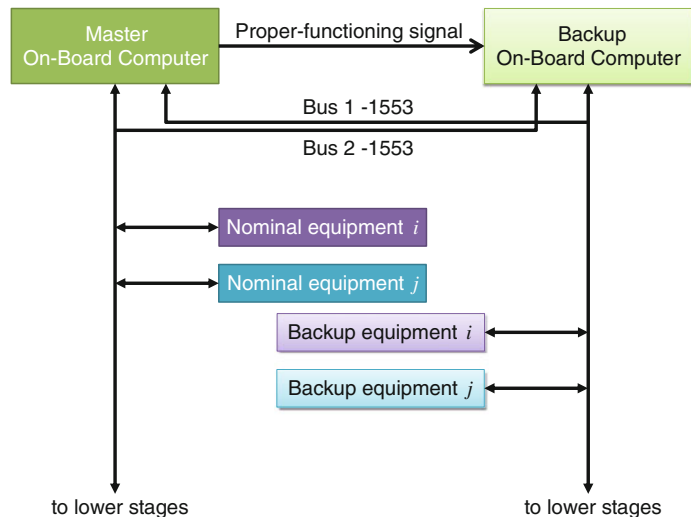
Avionics and software are obviously fundamental in the operation of a launcher, both during the ground phase and in flight. The main requirements applicable to launcher avionics and software are

- Implementation of the guidance, navigation, and control functions. This subsystem must trigger the propulsion, guide the launcher throughout the flight, control the launcher during all phases, jettison the stages and fairing, control the payload attitudes for orbit injection and safely deploy them, and perform the end of life actions.
- Implementation of the stages. This subsystem must take care of engine ignition/shutdown, and perform operational monitoring of the propulsion system, and the pressurization and propellant levels of the tanks. It also participates in the thermal control functions of the stages, and commands their passivation at the end of the mission.
- Avionics and software is in charge of the neutralization function.

The avionics and software subsystem must also integrate numerous constraints

- It must tolerate the natural and launcher-induced environments: namely this dynamic environment (sinusoidal vibrations, random vibrations, shock, transport), the local ambient environment in terms of temperature and humidity, and the effects of pressure/depressurisation, and the natural radiation environment (protons, heavy ions, etc.).
- It must cope with product assurance requirements: in particular, safety constraints that can be either qualitative, such as fail-safe/fail-safe (FS/FS) concepts, or quantitative with probabilities of success. The reliability requirements of the mission often dictate the level of redundancy, and the availability constraints influence the technological choices.

Fig. 8.14 Typical data flow architecture of launcher avionics equipment



The overall architecture of the avionics subsystem varies from one launcher to another. It is often centralized in a dedicated avionics bay, with a limited amount of equipment distributed along the launcher, near the elements they are serving. Some equipment may also be located on lower stages depending on the mission profile, for example, when they cease to function early in the flight.

The electrical equipment is often structured around a data bus, with norms such as MIL-STD-1553B; see [Chap. 15](#). A typical architecture is presented in [Fig. 8.14](#).

8.7.1 Launcher Electrical Power

The purpose of the electrical power subsystem is to supply the necessary power at a roughly constant voltage, satisfying the need for high power when required and enabling ground to on-board switching.

The subsystem's architecture includes various electro-chemical power sources and terminal equipment connected to sources via appropriate distribution equipment (power distribution unit, safety unit).

Power sources consist of primary sources, which are non-rechargeable voltaic-cell batteries, and secondary sources, which are rechargeable accumulator batteries. Three types of electro-chemical combinations are used: nickel-cadmium is a low cost, proven technology; silver-zinc is an attractive specific power and energy source, and lithium-based combinations yield a high level of specific energy. More details can be found in [Chap. 10](#).

8.7.2 Safety, Telecommand

The safety function aims to protect personnel at the launch base, and populations of the overflow territories in the

event of a problem that could lead to a critical failure such as a trajectory deviation. In this case, a neutralization order can either be sent from the ground, elaborated by the launcher on loss of stage integrity, or by the on-board software algorithms. Neutralization is achieved by fragmenting the launcher, or by simply cutting the propulsion in order to drop it into a safe zone.

The safety subsystem fulfills three vital functions: automatic and instantaneous destruction or engine shutdown in accordance with on-board criteria; destruction or engine shutdown, on reception of a command from ground control; and automatic delayed destruction after a nominal separation in order to avoid leaving floating wreckage.

The equipment necessary to perform these functions usually includes

- A dedicated battery to provide electrical power.
- A safety unit for distributing power (including ground to on-board switching), dealing with the separation mechanisms, processing the orders for destruction, and actuating the devices that perform the destruction.
- A radar transponder, with several aerials for trajectory determination.
- A destruction telecommand receiver, with its own set of aerials.

8.7.3 Flight Control

The flight control subsystem is in charge of performing all the guidance, navigation, and control (GNC) activities previously identified. These functions are detailed in general form in [Chap. 12](#), but includes dedicated sensors, among which are the inertial measurement unit (IMU), the gyrometric block (BGY) and other measurements used to generate the inputs to the GNC loop.

- The aim of the three-axis IMU is to provide attitude angles and cumulated accelerations. It is often strapped down, based on three-axis gyrolasers and pendular accelerometers.
- The BGY is used, when necessary, to provide angular velocities in yaw and pitch for the control of some of the low frequency modes. It can be based on fiber optic technologies.
- A dedicated functional unit acquires such pressures, temperatures, propellant and pressurant levels, and other sensors used in the operational algorithms.

The outputs are the orders that are sent to the engine actuators and to the servo-control mechanisms that operate the flight controls, including the attitude control system. It also issues the sequential commands to advance the launcher through its various transitory phases. These functions are performed within the on-board computer.

8.7.4 Telemetry

The function of the telemetry subsystem is the acquisition and processing of all the on-board flight measurements, their conditioning, their monitoring, and all activities in preparation for transmission to ground stations: data recording and retrieval, formatting, modulation and transmission through aerials located on the external surface of the launch vehicle.

8.8 Future

8.8.1 Systems

The future of space launch systems will be dominated by the evolution of existing systems. Expendable multi-stage rocket propelled launch vehicles are well suited to currently envisioned transportation demands. There are, however, opportunities that might lead to a so-called rupture or breakthrough revolution, or at least an accelerated evolution.

- Technology evolution (high-energy density propellants, nanotube structures, etc.).
- An essential increase of institutional and/or commercial space launch demand (defence against natural and human-induced hazards, orbital tourism, etc.).
- Suborbital tourism launch systems and related operation experience allowing for an efficient entry into reusable orbital launch systems.

The evolution of present systems will mainly be influenced by

- The demand for launcher evolution and the related size, mass, and other features of payloads.
- Institutional budgets for technology improvement, demonstration, and system development.
- International cooperation.

It is also noteworthy that the overall goal of reliability, efficiency, and finally cost improvement is intrinsic for the involved industry.

With the retirement of the Space Shuttle in 2011, systems for human space launch will stay even closer to the existing launch systems. The human ‘payload’ requires a specific compartment or capsule to provide life support during the entire mission, including reentry. Delta qualification and additional emergency rescue systems should allow most of the current mid-to-heavy payload launch systems to be upgraded to human-rated systems as and when required.

8.8.2 Research and Technology

Institutional technology programs and industry research and technology funding are the major sources for technology evolution. Due to the specific constraints for space launch systems, technology spin-in and spin-off has been and will remain limited to basics and components. Due to the importance and complexity, a major effort will have to be concentrated on the propulsion subsystem and related components without missing a balanced evolution of all related technologies. Some obvious focal points can be identified, but should always be oriented towards a concrete system development and related requirements: high energy density chemical propulsion, chemical (cryogenic) liquid propellant rocket engines reusability, lightweight structures for low and high temperature, and health monitoring for components and systems.

8.8.3 Demonstrators

New technologies and system configurations, or entirely new systems, present additional risks of failure. Demonstration or qualification flights are the essential measure of risk mitigation prior to entering the operational phase. Depending on the leap in technology and the system configuration, there might be the need for specific (mostly subscale) demonstration vehicles. There are some focal points in technology and system configuration evolution for future and advanced space launch systems crying out for specific demonstration vehicles and missions: high speed

(>Mach 3) atmospheric flight (ascent and reentry), application of air-breathing propulsion, and application of tethers.

Of course, in-flight demonstrations should only be applied in cases where on-ground demonstration is either not (efficiently) possible or not sufficiently representative of environmental conditions. Even then, budget constraints will often prevent the efficient use of demonstration opportunities within the development of future space launch systems.

8.8.4 Advanced Concepts

Advanced concepts for space launch systems include a wide variety of options and ideas that are driven by the following parameters

- Part or full reusability (e.g. for a multiplicity of missions expendable upper stages are necessary).
- Number of stages.
- Launch and landing method (horizontal, vertical, with/without propulsion, lifting body or winged).
- Propulsion (rocket, air-breathing propulsion, combinations).

Experience gained from 30 years of Space Shuttle operation and a great many national and international technology and demonstration programs has provided an essential understanding of the evolution opportunities for and the constraints imposed upon advanced concepts. That is

- Single-stage vehicles will only become feasible after essential progress in propulsion and structures technologies.
- Reuse of first or boost stages leads to limited cost savings that do not justify the development, infrastructure, and operational expenditure.
- High-speed air-breathing propulsion is very complex and its integration will lead to even more complex system configurations for which an enormous demonstration and development effort will be necessary.
- Horizontal unpowered landing with wings is the least-effort solution for the return of large rocket stages.

The analysis of advanced concepts is done mainly by numerical simulation. The necessary software tools should represent the available knowledge, and therefore integrate knowledge from a variety of typically geographically dispersed specialists via a network. NASA has reached a very high level of simulation and optimization tools. Maintaining this level and stepwise improvement and verification of the simulation algorithms might also be a challenge for the future.

Further Reading

1. Sutton, G.P. and Biblarz, O., "Rocket Propulsion Elements", Eighth Edition, Wiley, 2010.

Gerard Miglioreno and Torben K. Henriksen

9.1 Space Vehicle Structures

Structural elements form the backbone of a spacecraft. They provide the overall mechanical integrity of the spacecraft under launch and in-orbit loads. Furthermore, the structure ensures that the spacecraft configuration is maintained during all mission phases, ensuring the relative alignment of components like antennas, reflectors, sensors and optical instruments. In the latter cases, the potential loss of dimensional stability of the spacecraft structure can severely degrade the mission performances.

Spacecraft structures typically consist of the following structural elements: shells of revolution (e.g. cylinders and cones tube), ‘sandwich’ panels (e.g. equipment and instrument platforms), rings, bars, and trusses. An exploded view of a typical spacecraft structure is illustrated in Fig. 9.1.

The spacecraft structural architecture is driven by the spacecraft attitude control requirements. Spin-stabilized spacecraft have a rotationally symmetric structure, whereas ring-type equipment platforms are arranged around a central cone or cylinder. Three-axis stabilized spacecraft are usually a box-type structure with large external panels providing sufficient space for the accommodation of instrumentation, antennas and a central tube load-carrying structure with flat sandwich panels attached.

Spacecraft structures are typically divided into primary and secondary structures. The primary structure defines the main load path down to the base of the spacecraft where the spacecraft interfaces connect with the launch vehicle. The primary structure determines whether the mechanical design is compliant with the mechanical requirements of the launcher, such as the minimum natural frequency (stiffness)

requirement and quasi-static load cases (strength). Typical primary structure design components are thrust tubes (cylinder, cone), the launch vehicle interface ring, sandwich panels and struts to support the panels, tanks, and other spacecraft equipment.

The secondary structure provides mounting provisions for the payload, units, solar arrays, antennas, etc. It transfers load to the primary structure. Typical secondary structure components are sandwich panels. Attached to the spacecraft are deployable flexible appendages such as antenna reflectors and solar panels.

Due to launcher and mission constraints, spacecraft mass is a key design parameter in the design of all spacecraft. The mass of a spacecraft structure will typically not exceed 10–15 % of the spacecraft dry mass [1–3].

Material selection is important in the design of a spacecraft structure. For example, high specific strength and stiffness can be important material performance parameters in achieving strength and stiffness requirements in a lightweight design; however, the material parameters will be a function of the failure mechanism being considered, as such a range of efficiency parameters can be defined. Composite material have high specific strength and stiffness properties and therefore are candidate materials for spacecraft structure applications. They possess more favorable stiffness-to-weight ratios than the conventional metallic materials (e.g. aluminum alloys). Furthermore, composite structure performance can be tailored to specific needs by choosing the structure layup accordingly; for example, by embedding high strength or stiffness carbon fibers, depending upon whether strength or stiffness requirements drive the design.

Where high dimensional stability under thermal loads is demanded, composite structures are superior to metallic structures due to their very small coefficients of thermal expansion (CTE), which might even become negative depending on the laminate layup. An additional means of achieving dimensional stability is the use of ceramics, which can be used for, say, optical benches or for telescope

G. Miglioreno (✉) · T. K. Henriksen
European Space Research and Technology Centre, European
Space Agency (ESA-ESTEC), Noordwijk, Netherlands
e-mail: Gerard.Migliorero@esa.int

Fig. 9.1 Exploded view of spacecraft structure showing base/interface ring

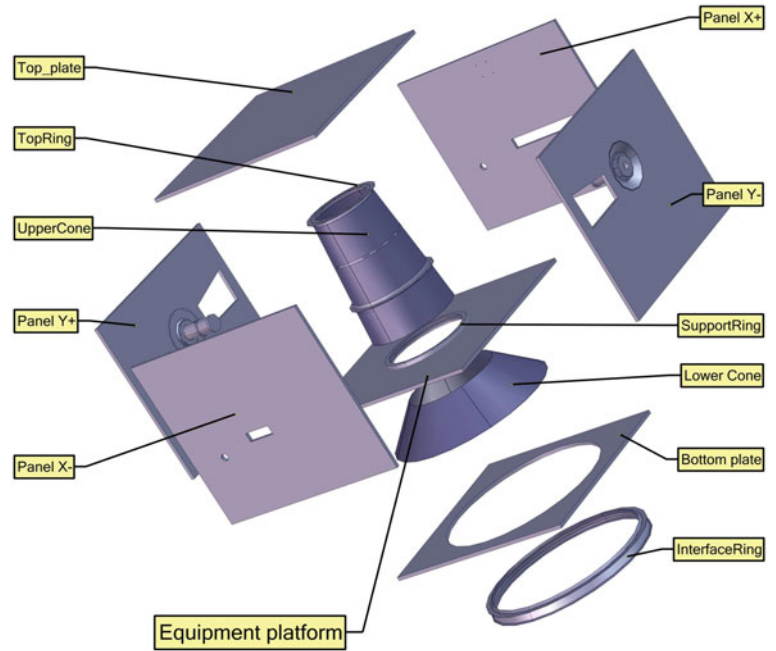


Fig. 9.2 Herschel Telescope.
Image EADS Astrium

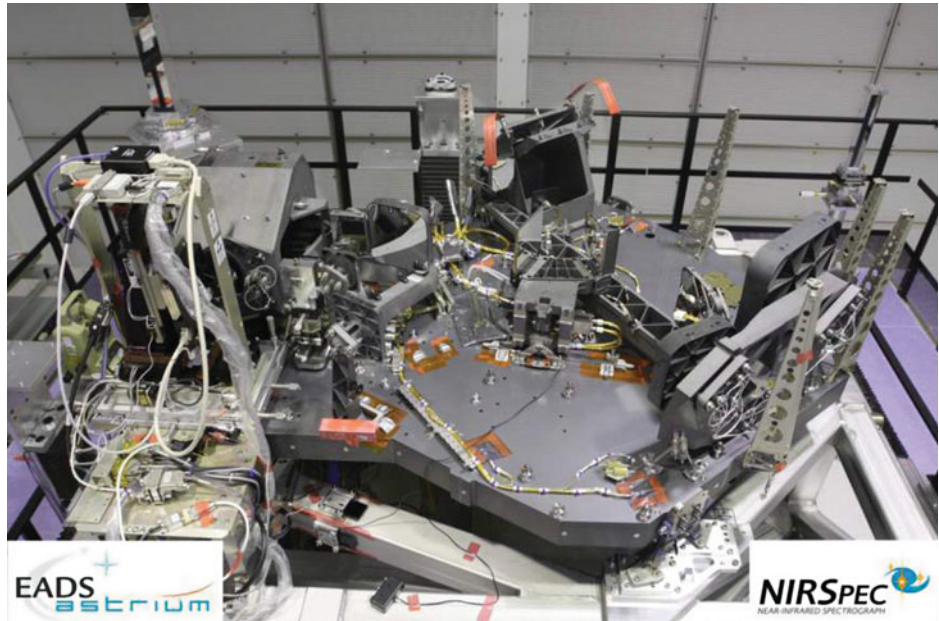


structures. Typical examples have been implemented in the Herschel-Planck Telescopes (see Fig. 9.2) and the Near-Infrared Spectrometer for James Webb Space Telescope (JWST) SiC100 optical bench (Fig. 9.3).

An important spacecraft structural subsystem is the separation system used to attach and separate the spacecraft to and from the launch vehicle, for example, by means of a clamp band. The main requirements the clamp band include withstanding the bending moment, running loads, and shear

loads during launch. In addition, the loads of the vibration tests on the ground and the transportation loads for the other specific separation systems used during the spacecraft development must also be endured. The clamp band must also ensure that the separation between the spacecraft and the launch vehicle occurs at the required separation velocity (0.5 m/s is used for most launch vehicles) and in the required attitude domains for the spacecraft and the last stage of the launcher. Finally, the clamp band should ensure

Fig. 9.3 The near Infrared Spectrometer for JWST SiC100 optical bench and some optics. It operates at 30 K. *Image EADS Astrium*



a separation without debris and pollution in order to conform with any spacecraft cleanliness specifications. Separation is accomplished by actuating several bolt cutters to release the band so that torsion and pushers springs can push aside the clamps and ensure separation between the spacecraft and the launcher.

The major design/mission structural requirements for a spacecraft structure are discussed in detail in Sect. 9.2 of this chapter. Most of the spacecraft mechanical design requirements are specified in the launch vehicle user's manual.

9.1.1 Crewed Space Vehicles

When a human crew is part of a space mission, their safety becomes one of the main objectives of the mission itself. Therefore, for crewed spacecraft, safety requirements drive most of the spacecraft design.

Of course, the structure of a crewed spacecraft is one of the subsystems that are heavily involved in protecting the life of the crew, because of its functions of isolation, protection and support of vital equipment. A failure of the crewed vehicle structures will most likely result in a catastrophic failure leading to permanent injury or loss of life.

A crewed space structure is designed to fulfill various functions. The basic one is to protect the habitable crew environment from outer space. Additionally the structure is designed to provide and maintain a 'shirt-sleeve' environment (see Fig. 9.4), and to minimize the vibrations and acoustic pressure and shock loads transmitted to the crew and to critical systems, in particular during launch and reentry. It must constrain as well the deformation of human body below sustainable limits.

For long exposure to the space environment (for instance, the International Space Station) the structure must protect the crew against micrometeoroid/debris impacts throughout its operational life. The function of the crewed space vehicle structure is not only to protect the crew, but also all systems that are critical for the sustainability and protection of life such as life support, power generation, thermal control, communications, food and waste management, and aerodynamics.

The crewed space structure is designed to perform its functions during the complete mission, taking into consideration the possible material degradation due to exposure to different environments during its entire lifetime, in particular space environment exposure and atmospheric reentry. It must be designed to prevent or reduce other hazards such as pressurized system explosion, high touch temperatures, sharp edges, and the presence of toxic materials.

It must be noted that facilities, payloads, instruments, experiments and other systems that are to be operated inside a crewed vehicle must also be considered as well to be crewed spacecraft structures.

The design of crewed spacecraft structures must also take into account human factors and accessibility requirements in order to simplify not only nominal operations but also the means of access to the spacecraft and the recovery of the crew after landing. It can be required also to provide visibility of the outer environment (e.g. viewports, see the Cupola Fig. 9.5). In addition, structures and mechanism must not create noise and vibration levels that exceed the limits imposed by crew comfort and equipment constraints.

Given that safety is one of the main aspects to be optimized, a crewed structure has to overcome all the potentially hazardous functions not only in nominal mission

Fig. 9.4 European Space Agency astronaut Andre Kuipers, Expedition 30 flight engineer, prepares to insert ESA Role of Apoptosis in Lymphocyte Depression 2 (ROALD-2) experiment samples into a Minus Eighty Laboratory Freezer for ISS (MELFI-1) dewar tray located in the International Space Station's Kibo laboratory. *Image* NASA/ESA; ISS030-E-033272 (24 December 2011)



Fig. 9.5 Full panoramic view of Earth from the Cupola of the International Space Station. *Image* NASA; ISS022-E-066963 (17 February 2010)



scenarios, but cover all credible contingencies. The application of damage tolerance or fault tolerance principles for the mitigation of catastrophic and critical hazards is an essential part of the structural verification of space flight vehicles, modules, payloads, and ground support equipment. Mechanical failures that may result in loss of life, severe injury, or major damage to the hardware, must be prevented. For that purpose, damage tolerance principles are applied to ensure that undetected cracks and other defects existing in the structure do not lead to failure within the service life of critical hardware. All structural items whose failure would result in a catastrophic or critical

hazard (e.g. disintegration, loose item, crew impact, loss of critical function, jamming, et cetera) are designed and verified according to fracture control principles. Redundancy is the preferred solution to reduce the risk that the failure of a structural element having catastrophic consequences. When the implementation of redundant structures is not feasible, the design shall rely on a damage tolerance application to ensure that no defect will grow into a complete failure in the design lifetime.

The Columbus laboratory module is shown in Fig. 9.6 permanently attached to the International Space Station (ISS). It is a cylindrical structure built from welded

Fig. 9.6 S122E009992 (18 February 2008) A close-up view of the Columbus laboratory module (*center*), permanently attached to the ISS. *Image* NASA; S122-E-009992 (18 February 2008)



Fig. 9.7 S128E007203 (1 September 2009) The SOLAR payload; external payload of the ISS is seen center of image as astronauts John Olivas and Nicole Stott (*right*), both STS-128 mission specialists, participate in the mission's first session of extravehicular activity (EVA) as part of the construction and maintenance of the Station. *Image* NASA; S128-E-007203 (1 September 2009)

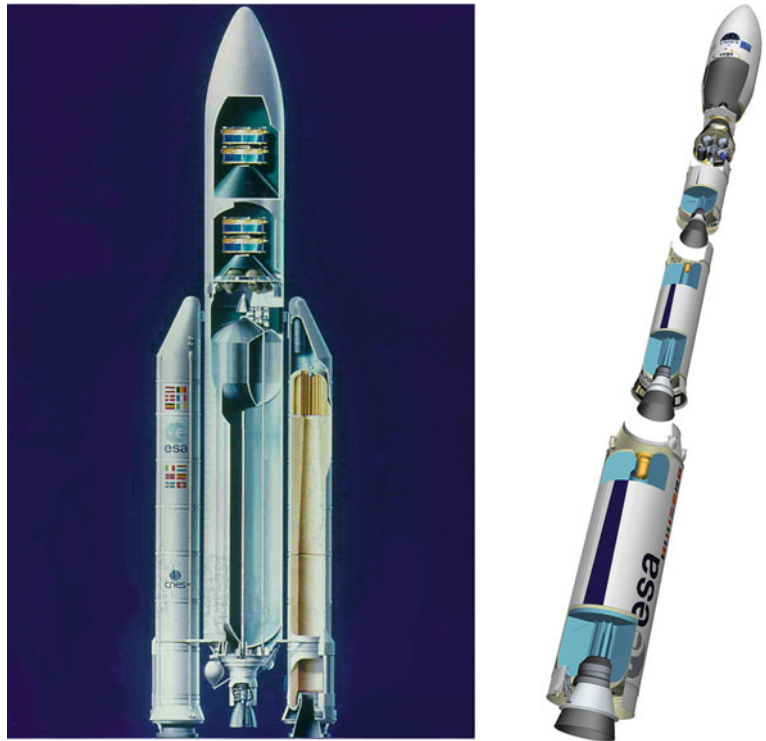


aluminum panels (Al 2219). The secondary structures (stand-offs and racks) are built from aluminum as well. The meteoroid and debris protection panels are built in aluminum and Kevlar. Note that crewed space structures are mainly built of metallic materials, especially aluminum. The SOLAR payload is shown in Fig. 9.7 on the outside of the Columbus module. Its structure was built of aluminum and carbon-fiber sandwich panels in order to minimize Sun-pointing distortions.

A typical crewed vehicle has generally to withstand all the loads that are common for robotic vehicles. Therefore, according to the specific launcher requirements, low

frequency accelerations, random vibration, acoustic pressure and shocks, are generally applicable load-cases to be considered. In addition, crewed structures are designed to sustain pressure loads and crew loads (handling loads but also inadvertent ones such as kicks). Pressure loads can vary from about 1 bar of change for hulls to hundreds of bars for pressure vessels. Crew loads can be as high as 560 N for inadvertent kicks, applied on a relatively small area. In addition, when a crewmember has to operate a mechanism, the reaction forces must be low enough to allow the intended crewed actuation. Both pressure loads and kick loads can be critical for shatterable materials that, if they are

Fig. 9.8 Ariane 5 with the Cluster spacecraft and Vega launch vehicles. *Image ESA*



not contained and therefore physically separated from the crew, must not be allowed to break.

For some crewed structures it is also necessary to accommodate the high loads experienced during reentry into the atmosphere and the subsequent landing. The verification campaign can be quite intensive, and the design loads are sometimes driven by test loads (both for qualification and acceptance). In particular, proof tests for pressurized systems are usually more severe than other loads experienced during the mission. Also in case of the proto-flight approach, such as without development of a prototype qualification model, the structure is designed to withstand qualification loads as part of its design life.

9.1.2 Launch Vehicle

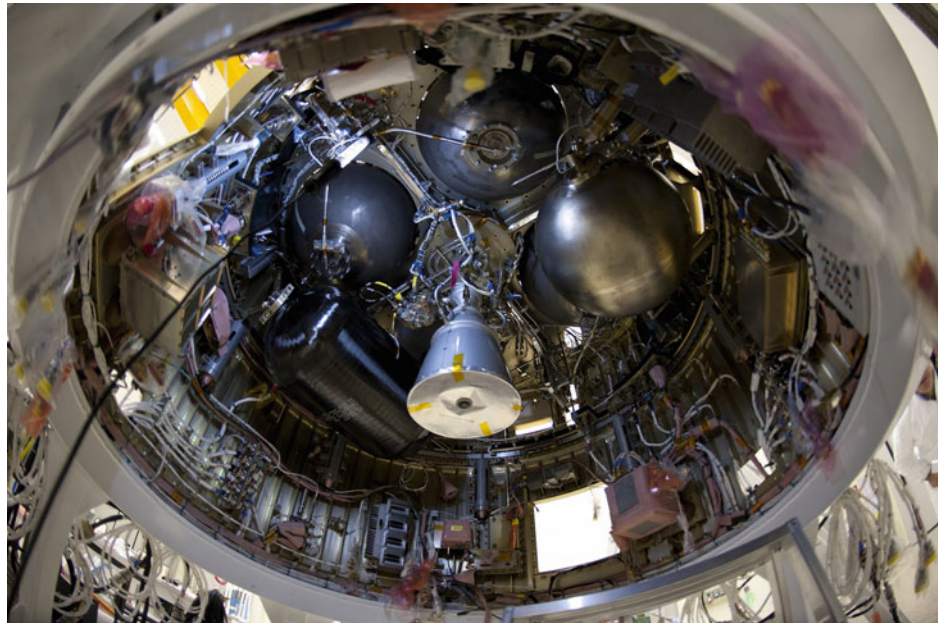
Launch vehicles such as Ariane 5 and Vega can be characterized as respectively multi- and mono-body launchers; refer to Fig. 9.8. The bodies of the launchers are slender, tall structures designed to withstand loads such as longitudinal acceleration as well as the lateral loads due to blast waves, atmospheric gusts, and thrust vector control steering. In addition, the substructures such as cryogenic stages, boosters, and solid rocket motor stages, are loaded by internal pressurization ranging from a few bars to approximately 100 bars of differential. The stages of the launcher are each jettisoned after burnout in order to facilitate efficient transportation of the payload into low Earth orbit. This

may also induce engine cut-off loads and subsequently shock loads due to stage separation.

In terms of architecture, launch vehicles can be broken down into inter-stages, solid rocket motor boosters (Ariane 5), solid rocket motors (Vega), cryogenic propulsion stage (Ariane 5), upper composite liquid propulsion stage, and a payload fairing. The Vega substructure architecture is described now as an example. The main parts of the cryogenic and solid propellant propulsion stages are pressure vessel tanks made of aluminum, steel or composite overwrapped pressure vessels (COPVs) which have to sustain loads induced by internal pressure as well as flight loads. The dome structures at the forward and aft ends of a tank have to withstand the axial loads induced by the internal pressure; refer to Fig. 9.9.

In addition, the aft dome of Vega accommodates the nozzle loads via the flex joint that attaches the nozzle to the polar boss of the aft dome. The actuators of the thrust vectoring control system are attached on the one side to the aft frame of the motor skirt and to the other to the nozzle attachment points. The various cylindrical or conical inter-stages form the junctions between consecutive stages, and are split into two parts by pyrotechnic devices at stage separation. The liquid propulsion stage of the upper composite houses the oxidizer and fuel propellant tanks (1st level) and the avionics for launcher attitude control (2nd level). The payload adapter, made of carbon-fiber-reinforced polymer, CFRP, composite, connects the payload to the launcher via a pyrotechnic clamp band device. The

Fig. 9.9 Vega's Attitude and Vernier Upper Module (AVUM) liquid propulsion stage. *Image* ESA



payload fairing is made of a CFRP sandwich shell and protects the payload from wind, rain and hail during atmospheric flight, as well as from acoustics and thermal loads.

The mechanical requirements for the design and development of a launch vehicle, in particular for structures, are different to those of space vehicle structures. They are prescribed for each subsystem (stage, inter-stage) separately. The loads are typically a mix of quasi-static (linear loads and bending moments) plus dynamic (sine, acoustic). For cryo-temperature stages, the thermal loads are of great importance and drive the design. Additionally, the pressure loads induced in the solid booster cases during ignition could also drive the design of these elements. The ECSS-E-STD-32C and other standards and handbooks mentioned in [Sects. 9.16.1](#) and [9.16.2](#) provide detailed requirements.

The limit loads are in fact the design limit loads and the factor of safety logic is presented later, see [Fig. 9.17](#). The safety factor KQ is typically 1.25 and the acceptance factor of safety $KA = 1.0$; other factors are dependent on the project. In order to decouple the launcher vehicle from the engines, pumps etc., certain stiffness requirements for the complete launch vehicle are required and translated into stage and inter-stage stiffness requirements.

The space mission environments for launch vehicle or spacecraft structures are associated to the same dimensioning load cases. The categories of environments include static, low frequency transient, low frequency harmonic, acoustic and shock environments. The static environment is associated with the acceleration of the launch vehicle along its flight trajectory, and applies to the vehicle center of mass. The low frequency dynamic environments induced by transient and harmonic low-frequency loads are combined

with the static environment to obtain the complete total low-frequency load environment. These environments are dimensioning loads for interface strength verification and the strength assessment of the load-carrying structures.

9.1.3 Reentry Vehicle

A reentry vehicle is the part of a spacecraft that reenters Earth's atmosphere. The best-known example is probably the US Space Shuttle (now retired). The most used reentry vehicle in history is the Soviet/Russian Soyuz descent module, first used in 1966; see [Fig. 9.10](#).

In 1985, Europe decided to develop its own capabilities for crewed space flight and the development of the Columbus laboratory module, the Ariane 5 heavy-lift launcher and the Hermes spaceplane followed. Initial study and pre-development work was concluded for Hermes in 1992. The program was stopped due to difficulties in attaining the cost or performance goals, but European efforts on the development of reentry vehicles continued with the X-38 and the Atmospheric Reentry Demonstrator (ARD). The X-38 was a pathfinder, in cooperation with NASA, towards the Crew Return Vehicle (CRV) of the International Space Station. The ARD was a demonstrator capsule, whose shape was based on Apollo. It successfully flew on board an Ariane 5 in October 1998. The latest European developments for reentry vehicles are the Inflatable Reentry and Descent Technology (IRDT) demonstrator, the European EXPERIMENTAL Reentry Testbed (EXPERT) and the Intermediate eXperimental Vehicle (IXV). EXPERT is illustrated in [Fig. 9.11](#), and IXV in [Fig. 9.12](#).

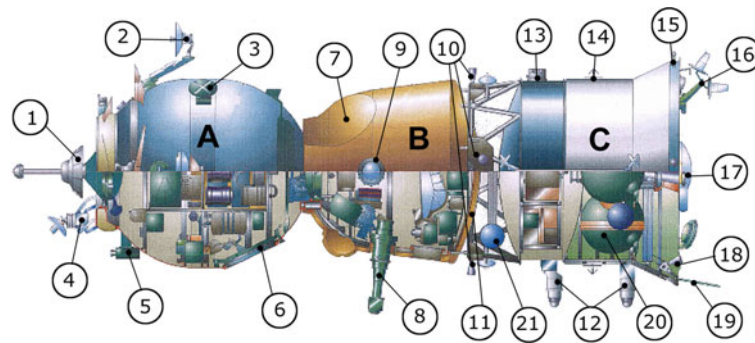


Fig. 9.10 Soyuz TMA Spacecraft diagram. Orbital module (A); 1 docking mechanism, 2 Kurs antenna, 3 television transmission antenna, 4 Kurs antenna, 5 camera, 6 hatch. Descent module (B); 7 parachute compartment, 8 periscope, 9 porthole, 11 heat shield.

Service module (C); 10 and 18 attitude control engines, 12 Earth sensors, 13 Sun sensor, 14 solar panel attachment point, 16 Kurs antenna, 15 thermal sensor, 17 main propulsion, 19 communication antenna, 20 fuel tanks, 21 oxygen tank. *Image NASA*

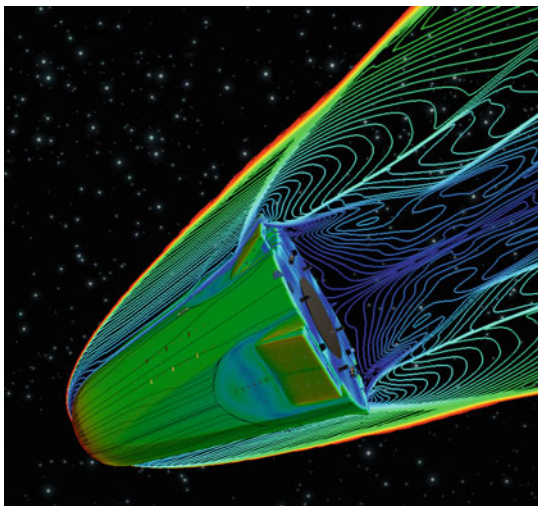


Fig. 9.11 A CFD analysis of the EXPERT vehicle during reentry. *Image DLR*

In the vehicles mentioned previously, different structural concepts were used for the primary structure. Winged vehicles such as the Space Shuttle, Hermes and the X-38 applied typical airplane construction, using frames and stringers in combination with a metallic skin. The ARD used a front shield structure and a lateral cone made of stiffened metallic panels. The demonstrators EXPERT and IXV have innovative configurations: a full hot-structure configuration for EXPERT and a monocoque configuration with several bulkheads for IXV. In the special case of inflatable reentry vehicles, part of the heat shield structure is made of a flexible envelope inflated by gas.

The primary structure is the structure that provides the external geometry of the vehicle, and sustains the external loads. The primary ('cold') structure is made of standard aerospace materials (such as aluminum alloys, CFRP, or even titanium alloys), and is protected by a thermal protection system (TPS). The TPS may be made of ablative

materials. However, in some cases it is more efficient to select a concept based on 'hot structures'. A hot structure provides the required strength and stiffness and is able to sustain the high temperatures. Typical examples are the vehicle nose, wing leading edges, and control surfaces (rudders and flaps). These are made of ceramic matrix composites. More details are given in Sect. 9.5.

Reentry vehicles are subjected to specific load cases during their mission, such as to ground, transportation and launch loads. These vehicles are also subjected to reentry, descent and landing loads. The entry interface is typically defined starting at 100 km altitude. This load case involves a heat flux applied on external surfaces of the vehicle due to aerodynamic heating; this heat flux is proportional to ρV^3 , where ρ is the atmosphere density and V the vehicle velocity. Whilst traveling through the denser layers of the atmosphere, the external surfaces are subjected to dynamic pressure that decelerates the vehicle; this deceleration is proportional to the dynamic pressure $\frac{1}{2}\rho V^2$. In case of active control surfaces (such as flaps), local loads are applied to the structure during reentry. These heating, deceleration, and local loads evolve during the reentry and do not have their maxima at the same time. It is necessary to define dimensioning load cases. The descent phase follows the reentry. During this phase, the vehicle is decelerated to the required end velocity. Aerodynamic decelerators, such as parachutes, are used. Opening a parachutes imposes a sudden deceleration (shock) on the vehicle. Finally, the vehicle is brought to rest on the Earth's surface. Different landing options are available: on land (Soyuz capsules), on water (Apollo capsules) or on a runway (Space Shuttle, or USAF X-37B). In all cases, the residual kinetic energy must be dissipated to limit the resulting loads on the vehicle. Various types of structures are available for this energy absorption: parachutes, landing gears with dampers, crushable structures or airbags (or a combination thereof). An airbag and a parachute landing system are shown in Fig. 9.13.

Fig. 9.12 The Intermediate eXperimental Vehicle under construction (*left*) and visualized during reentry (*right*)



Fig. 9.13 Two different landing options; Engineers test huge, multi-lobed air bags, the type which protected the Mars Pathfinder spacecraft before it impacted the surface of Mars, (*top*) and the Dragon spacecraft shortly after splashdown and being dragged by the main chutes (*bottom*). Image NASA (*top*) and US Navy (*bottom*)



Space environments are hostile for reentry vehicles because the structure is subjected to a combination of extreme environments. In particular, the vehicle is subjected to a plasma of dissociated species (oxygen, nitrogen, etc.), which

interacts with the external surfaces. When hot structures are made of ceramic matrix composites, a quick erosion of the surface through ‘active oxidation’ of the material is expected. During the descent in the lower part of the atmosphere, the

Fig. 9.14 Pressure vessel of the xenon tank of the Dawn probe.
Image NASA/JPL



properties of the surrounding air will affect the vehicle loading and external structures. In particular, winds and gusts will affect the descent, and meteorological phenomena (rain, hail, sand blast, salt air, etc.) may influence the strength and stiffness of structures. The surface conditions at landing are important parameters to define the landing loads: runway roughness, sea waves, rocks on ground, etc.

The structures and materials of reentry vehicles differ from structures of a typical spacecraft, due to the additional specific loads and environment to which they are subjected during entry, descent, and landing. The opportunity to return a space vehicle to Earth also opened a new perspective: reusability.

9.1.4 Pressurized Structures

A pressurized system is a system in which pressure loads are important for its design. An important standard for pressurized systems is ECSS-E-ST-32-02C. Typical cases on-board spacecraft are the tanks of the propulsion system, as shown in Fig. 9.14. Also liquid propulsion based launchers are essentially pressure systems, because fluid flows from the tanks to the nozzle, pressurized by stored high-pressure gas or pumps.

On crewed space vehicles, pressurized systems can have less conventional architectures, for example elements of life support systems, or containers used for biology or physiology experiments. Moreover, the nature of pressurized components differ, including human modules, vessels that have to contain fluids at hundreds of bars of pressure, piping, valves and bellows, and many others. Therefore, it is

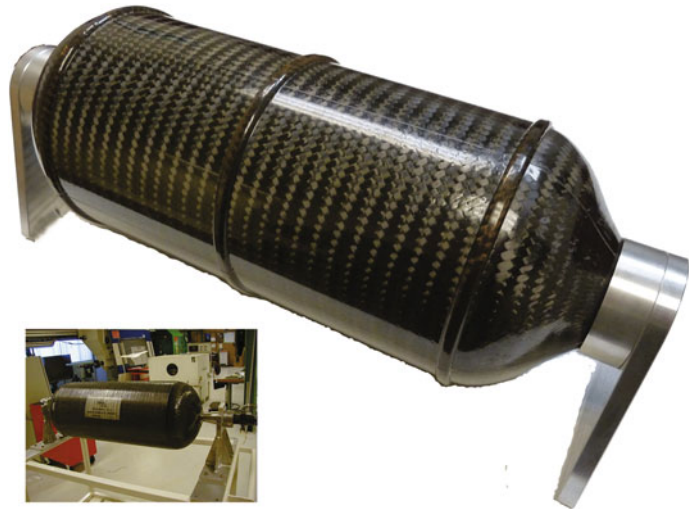
difficult to define a unique verification strategy valid for every pressurized system. Before any verification plan is established, a pressurized system has to be classified in terms of lower level categories. However, pressurized systems have in common the fact that their failure usually means the failure or degradation of the mission. This is true not only because of the importance of their functionality, but also because they store and distribute energy. If such energy is suddenly released it could endanger the integrity of the whole spacecraft or space vehicle.

An early step in verification is defining the limit design load, which is usually referred to as maximum design pressure, external mechanical static, dynamic and thermo-elastic loads. In human space flight or ground operations in the presence of humans, a high level of safety is required. For these pressurized structures, the maximum design pressure is that calculated by taking into account the one or two worst-case failures that could occur.

A further aspect to be considered in the design of pressurized systems is the compatibility of the container with its contents. Special care must be taken in systems containing highly reacting fluids. For example, only low flammable materials can be used in pure O₂ systems. When ignition is not desired and a flammable atmosphere could be present, the design must not include any ignition source. In addition, stress corrosion cracking (SCC) and hydrogen embrittlement are two phenomena that are of concern and should be avoided by appropriate design controls.

The first step in designing and verifying a pressurized system is the classification and definition of the applicable requirements. Pressurized hardware can be classified according the following categories

Fig. 9.15 A braided composite overwrapped pressure vessel and inset the 70l xenon COPV. Image A&P Technology (main) and EADS Astrium (inset)



- Pressure vessels
- Pressurized structures
- Pressurized components
- Special pressurized equipment.

Pressure vessels are pressurized hardware designed primarily for the storage of pressurized fluid with high energy and/or pressure level. The usually accepted limit for a container to be classified as a pressure vessel is an energy level greater than or equal to 19,310 Joules or a pressure greater than or equal to 0.69 MPa. This limit is based on the potential energy within a pressurized gas assuming that it will adiabatically expand. The formula for calculating this energy is

$$E = \frac{P_1 V}{\gamma - 1} \left[1 - \left(\frac{P_1}{P_2} \right)^{\gamma - \frac{1}{\gamma}} \right] \quad (9.1)$$

where E is the stored energy (J), P_1 and P_2 are the internal and external pressures, respectively (Pa), V is the pressurized volume (m^3) and γ is the ratio of specific heats for the gas.

A pressurized vessel commonly used in crewed and robotic spacecraft, as well as in launchers is the composite overwrapped pressure vessel (COPV). A COPV is a vessel consisting of a thin, non-structural liner, wrapped with a structural fiber composite designed to contain a fluid under pressure. The liner provides a barrier between the fluid and the composite, preventing leaks (which can occur through matrix micro-cracks that do not cause structural failure) and chemical degradation of the structure. The most commonly used composites are fiber-reinforced polymer (FRP) using carbon and Kevlar fibers. The primary advantage of a COPV, both for high and low pressure applications, is the mass saving compared to monolithic metallic vessels. The COPV shall be designed to show Leak Before Burst

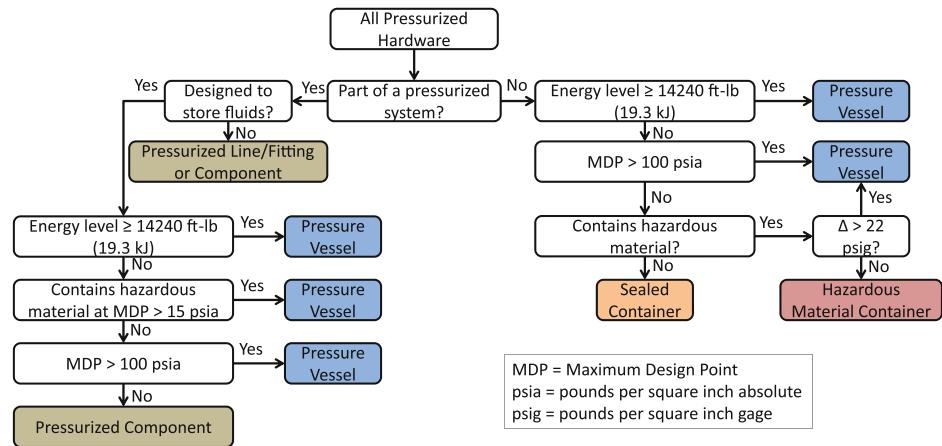
behavior when subjected to the Maximum Design Pressure (Fig. 9.15).

Pressurized structures are designed to carry both internal pressure and space vehicle structural loads, such as launch vehicle main propellant tanks, crew cabins and crew modules. A pressurized component is a pressurized system, other than a pressure vessel, pressurized structure, or special pressurized equipment, that is designed for the internal pressure. Valves, bellows, and connectors are examples of pressurized components. Special pressurized equipment with no straightforward classification includes heat pipes, cryostats, and hazardous fluid containers. Specific requirements have to be defined for each member of this category. Ultimate strength capabilities are usually verified by tests on qualification models (QM). Minimum burst factors commonly used are defined in the standards ECSS-E-STD-32-02C and ECSS-E-STD-32-10C, e.g.

- 1.5 for pressure vessels
- 2.0 for human modules and pressure vessels in human applications
- 3.0 for pressurised shatterable materials (e.g. glass windows)
- 4.0 for small lines and flexible lines.

Proof testing is done for acceptance of the pressurized flight hardware. The level of the proof test is designed to screen for defects that could grow and create failure during the complete lifetime of the structural item. Leak tightness tests and non-destructive inspections have to be performed after proof testing in order to verify the integrity of the system. A reduced verification program could be implemented for leaks before pressurized structures would burst, when the leakage does not result in hazardous consequences. Qualification and acceptance tests must take into account the operational environment of the structure to be verified. This is especially true for hardware that has to

Fig. 9.16 Typical classification of human space pressurized hardware



operate at temperatures that are very different from normal environmental temperature and also for pressurized structures that have to sustain additional loads besides the applied pressure.

The functionality of each component has to be verified during the acceptance campaign. In the design of pressurized systems, the designer must be aware that some components have a limited life (e.g. elastomeric materials used for O-rings) or need a recertification after some time (e.g. relief valves) (Fig. 9.16).

9.2 Mechanical Design/Mission Requirements of Space Vehicle Structures

The mechanical design requirements for a spacecraft (crewed and robotic), reentry vehicle, or pressurized structure are mainly imposed by the applied launch vehicle, mission, orbit, descent and landing. The mechanical design requirements for launch vehicle structures depend on a number of parameters, including the number of stages, boosters, payload mass, required orbit, launch trajectories, launch site, etc., and are specified during the development of the launch vehicle structures.

All space vehicle structures must withstand the launch environment, specified in the launch vehicle manual. The handling and transportation loads are, in general, lower than the test and launch loading conditions. However, attention must be paid to these loads. The most important mechanical design requirements mentioned in launch vehicle manuals are

- The spacecraft materials shall satisfy outgassing criteria, e.g. recovered mass loss (RML) and collected volatile condensable material (CVCN). Measurement procedure shall be in accordance to ECSS-Q-70-02A.

- The total rigid body mass and associated characteristic such as the center of mass and the second moments of mass of the space vehicle are restricted to prevent overloading of the launch vehicle payload adapters.
- The space vehicle shall meet minimum stiffness requirements both in launch and lateral directions in a stowed configuration. This is mostly manifested in minimum natural frequencies under specified boundary conditions (most times fixed at the interface between the space vehicle and the launch vehicle). The minimum natural frequencies depend on the total launch mass of the space vehicle. Minimum natural frequency requirements are posed to prevent dynamic coupling between the launch vehicle and the space vehicle. Therefore, the minimum natural frequency requirement is very important.
- The dimensioning of the space vehicle structure should be done using the specified quasi-static loads (QSL), such that the specified minimum stiffness requirements are met. This activity is more or less an interactive process. The quasi-static loads are typically specified as the combination of steady-state accelerations and low-frequency vibrations, which have no direct dynamic coupling with the space vehicle. The running load in the payload adapter between the launch vehicle and the space vehicle shall be as uniform as possible in order to prevent overloading of the payload adapter (line-load peaking). The stiffness, stability and strength of the space vehicle structure shall be verified by a static-load test. The QSL are increased by a test factor.
- Besides the quasi-static loads, dynamic mechanical, acoustic, and shock test loads are specified. The space vehicle shall survive these dynamic test loads, increased by test factors specified by the launch vehicle authority. This means that the test loads are more severe than the

real launch loads. The following mechanical dynamic loads are specified

- Sinusoidal enforced acceleration at the base of the space vehicle, normally in the frequency band between 5 and 100 Hz. The frequency range will be passed with a certain sweep rate in order to prevent dwell situations.
- Acoustical reverberant loads simulated in a reverberant acoustic chamber. This is idealization of the real launch acoustic loads is very difficult. The frequency range is between 20 and 10,000 Hz.
- Random enforced acceleration specification at the interface of the space vehicle. The random vibrations are not always specified. However, for instruments, equipment, tanks, etc., these random vibrations are of great importance. Mechanical random vibrations are due to structure-borne sound. The random vibrations are specified in a frequency range between 20 and 2,000 Hz.
- Shock loads are specified at the base of the space vehicle. The shock loads are specified by a shock response spectrum (SRS) in the frequency range between 100 and 10,000 Hz. The shock load environment is dependent on both the total mass of the space vehicle and the connection between the payload and the launch vehicle (clamp band or pyro-techniques).
- Depressurization during launch is considered in order to prevent the build-up of pressure loads in instrumentation and structures; for example, sandwich construction. Depressurization is prevented by implementing venting holes of adequate size.
- Mechanical interfaces with the launch vehicle, such as the dynamic envelope to prevent the payload striking the fairing. The interface is with the payload adapter through the clamp band.
- When in operation, the space vehicle shall meet stability requirements in order to avoid for example jitter at the line of sight due to micro-vibrations and thermally induced distortions.

In addition to the above mechanical design and test requirements, the following points may contribute to the understanding of the mechanical requirements

- The structural materials shall be selected in accordance with the ECSS standards, in particular ECSS-ST-032C and ECSS-E-ST-32-08C. A number of handbooks can be used for further evaluation of structural aspects.
- The mass properties are of great importance, and shall be evaluated during the design and manufacturing phases of the space vehicle (structures) project. Special tools or mass generators within the finite element analysis codes can be applied for monitoring the evaluation of the mass

properties. In the early stage, contingencies shall be considered in terms of maturity. Later in the project, such contingencies shall be removed.

- The design of the space vehicle structure is based on the specified quasi-static loads in combination with appropriate factors of safety (ECSS-E-ST-32-10C). Besides quasi-static loads, the stiffness of the space vehicle and dynamic loads are also important. The balance between the strength of the structure and its stiffness and dynamic load carrying capability is an iterative process. Undamped natural frequency properties of structural elements can be found in [4]. Informative books about the strength of material are [5–7]. For stability (buckling) analysis of structural elements the following handbooks are very useful: NASA SP-8007, 8019 and ECSS-E-HB-32-24A. Strength is the capability of the structure to sustain the design load without failure preventing mission success. This basic structural requirement for all types of space vehicles applies to all life cycle load events that the structure will encounter from manufacturing, assembly, ground handling, transportation and testing, launch and operation in flight. Insufficient strength leads to failure.

Static-stiffness and dynamic behavior of the space vehicle structure can be obtained from simplified models by making ‘hand calculations’ using software tools such as MATLAB[®], MATHCAD[®], Wxmaxima[®] and from more complex mathematical models using more advanced numerical tools such as the finite element analysis (FEA) method [8, 9], the boundary element analysis (BEA) method [10], and the Statistical Energy Analysis (SEA) method [11].

Damping in structures plays an important role in dynamic response (harmonic) analysis in the frequency domain. It is crucial to estimate damping characteristics from former projects. Space vehicles demonstrate, in general, low damping. Modal viscous damping models are frequently applied [12].

- The dynamic envelope is the physical space that the complete space vehicle must not exceed, while deflecting under static and dynamic loads to avoid contact between the space vehicle and the fairing of the launch vehicle.
- In orbit, the structural stability of the space vehicle is of importance because the structure must have the ability to maintain the alignment of the instruments, sensors and actuators mounted to it. This must be ensured after having survived the launch environment and during operations in orbit. Typical concerns are dynamic disturbers (e.g. reaction wheels, coolers), thermo-elastic distortions, permanent deformation, and slippage of mechanical connections. Structural stability must ensure that critical instruments, such as antennas, pointing devices and sensors, stay aligned in order to prevent performance degradation.

Table 9.1 Test factors, rate and duration (from the Ariane 5 User's manual)

Tests	Qualification		Proto-flight		Acceptance	
	Factors	Duration/Rate	Factors	Duration/Rate	Factors	Duration/Rate
Static (QSL)	1.25	N/A	1.25	N/A	N/A	N/A
Sine vibrations	1.25	2 oct/min	1.25	4 oct/min	1.0	4 oct/min
Acoustics	+3 dB (or 2)	120 s	+3 dB (or 2)	60 s	1.0	60 s
Shock	+3 dB (or 1.41)	N/A	+3 dB (or 1.41)	N/A	N/A	N/A

The loading environments encountered by the space vehicle during all phases define the design of the structure. In general, the load events are the ground activities (assembly, testing, transportation), the launch event, and the vehicle operations in space. When applicable, other important design drivers are in-orbit performance of spacecraft structures such as pointing accuracy and structural stability. Activities related to space mission environments, and mechanical loads are identified

- Space vehicle structures should not only survive the launch environment but protect the spacecraft non-structural components against the hostile space environment.
- The selected materials, non-structural as well as structural, must not unjustifiably degrade before and during the mission.
- Ground testing activities need to emulate mission environments with certain margins. As a result, test environments need to be defined early on, and structures need to be designed adequately to generate those environments.
- Mechanical loads have both static and dynamic components, and these are defined in launch vehicle user's manuals.

The design of a spacecraft vehicle structure and its subsystems must sustain all loads that it will experience. Typical loading events are listed in detail in standard ECSS-E-ST-32C, and summarized here

- Ground handling and transportation
- Liftoff
- Stage separations
- Stage ignition
- Stage or main engine cut-off
- Maximum aerodynamic pressure and gust conditions
- Spin-up and deployments
- Attitude control system firings
- On-orbit thermal environment
- Reentry
- Emergency landing (for reentry vehicles)
- When proto-flight, all of the ground tests designed to verify the above.

The above events induce steady-state, sine, and random vibrations, transients, and shocks. In special cases, temperature gradients may be expected.

The launcher authority will issue guidelines for the design and qualification of the spacecraft vehicle. These apply to the mounting interfaces at its base and are concerned with both quasi-static and dynamic loads. They will be denoted as flight limit loads (FLL), i.e., levels which are not to be exceeded with a probability 99 % and a confidence level of 95 %. The launcher authority will require the spacecraft designer to demonstrate that the design can withstand its qualification levels. Typical test load factors are given in Table 9.1, in which the distinctions between the qualification, proto-flight, and acceptance approaches is quite clear.

Flight acceptance testing is performed for space vehicle structures and equipment that have already passed test at the design qualification levels but where the workmanship remains to be tested. If a one-model program is followed (proto-flight approach, i.e., the prototype is actually flown), then the model must clear the higher qualification test; however, the duration and sweep rates are respectively lower and higher. It must be noted that in the proto-flight approach, full design qualification is only achieved when the mission is accomplished. In other words, the inherent risk of the proto-flight approach must be reduced as much as possible.

In the design phase, factors of safety must be applied; these are prescribed and provided in ECSS-E-ST-32-10C. The design logic and the application of factors of safety is illustrated in Fig. 9.17, where

- QL is the qualification load
- AL is the acceptance load
- DYL is the design yield load
- DUL is the design ultimate load.

The presented design logic in Fig. 9.17 is applicable for all types of space vehicle structures, where the coefficients are defined in Table 9.2. Typical factors of safety applied in the space vehicle design are presented in Table 9.3.

The typical factors of safety are denoted by

- KQ, the qualification test factor
- KA, the acceptance test factor
- FOSY, the yield design factor of safety
- FOSU, the ultimate design factor of safety.

Fig. 9.17 Logic of factors of safety application, ECSS-E-ST-32-10C

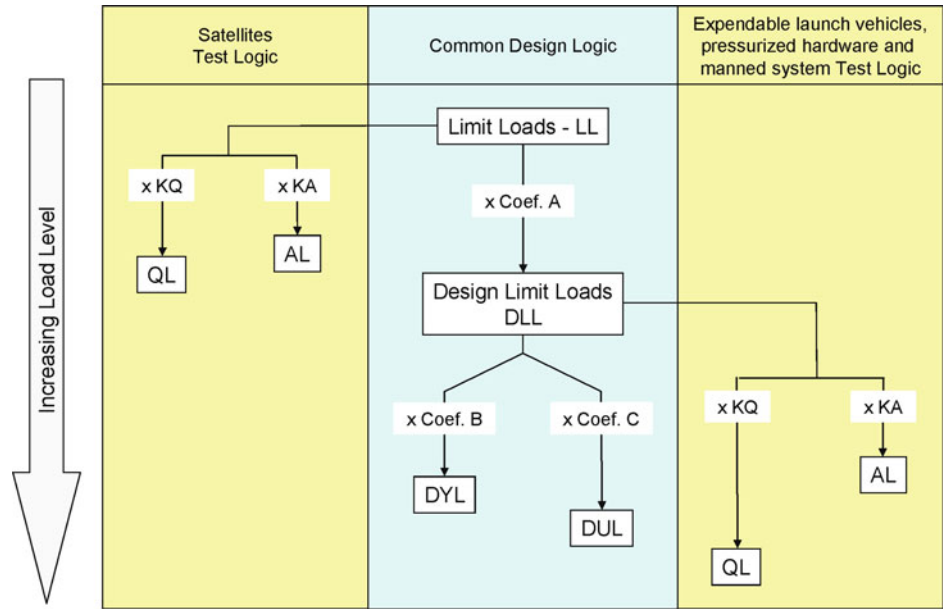


Table 9.2 Definition of coefficients in Fig. 9.17

	Satellite	Launch vehicle and pressurized hardware	Crewed system
Coef. A	$KQ \times K_P \times K_M$	$K_P \times K_M$	$K_P \times K_M$
Coef. B	$FOSY \times K_{LD}$	$FOSY \times K_{MP} \times K_{LD}$	$FOSY \times K_{LD}$
Coef. C	$FOSU \times K_{LD}$	$FOSU \times K_{MP} \times K_{LD}$	$FOSU \times K_{LD}$

Table 9.3 Typical factors of safety ECSS-E-ST-32-10C

Space vehicle	KQ	KA	FOSY	FOSU
Spacecraft	1.25	1.0	1.1	1.25
Crewed	1.4	1.2	1.25	1.5
Launch vehicle	1.25	1.0	1.1	1.25

It is required that at design yield loads (DYL), only non-permanent deformation occurs. This is only applicable for metallic structures. At design ultimate load (DUL) buckling will exceed allowable stress but there will be no failure of the structure. At DYL and DUL the margins of safety (MS) must be positive and sometimes above a certain value. The definition of the MS is

$$MS_y = \frac{\text{Allowable load (stress)}}{FOS \times \text{Design limit load (stress)}} - 1 \geq 0 \quad (9.2)$$

The space mission environments for launch vehicle and spacecraft structures are associated to the same dimensioning load cases. Different categories of environments exist such as static, low-frequency transient, low-frequency harmonic, acoustic, and shock environments. The static environment is associated with the acceleration of the launch vehicle along its flight trajectory and applies to the launch vehicle center of mass. The low frequency dynamic

environments induced by transient and harmonic low-frequency loads must be combined with the static environment to obtain the complete total low-frequency load environment. These environments are dimensioning loads for interface strength verification and the strength assessment of the load carrying structures. Shock and acoustic loads are usually dimensioning loads for optical equipment and appendages such as stowed solar arrays, reflectors and telescopes.

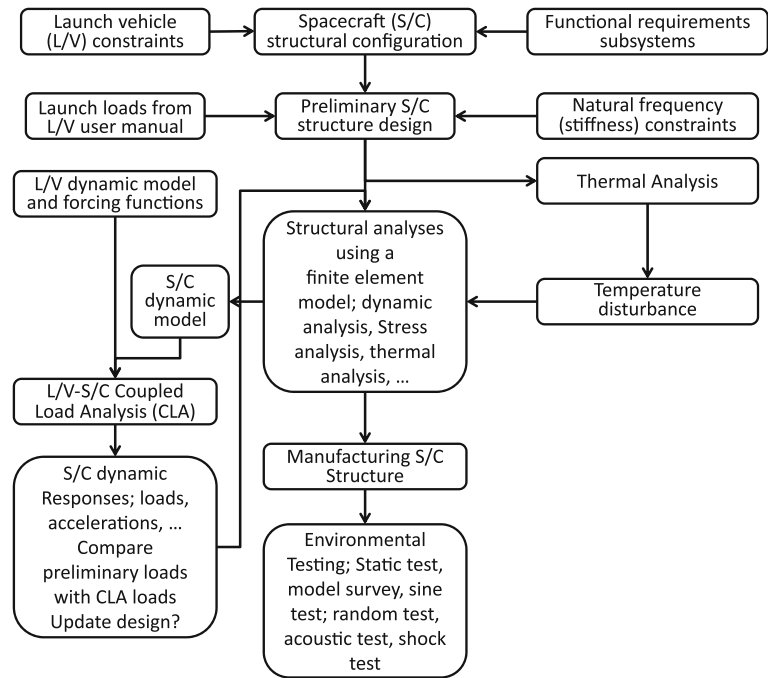
9.3 Verification of Space Vehicle Structures

In this section, verification by analysis and testing to qualify the space vehicle structure will be discussed. The important standards are ECSS-E-ST-10-02C and ECSS-E-ST-10-03C. The most common structural design validation cycle is shown in Fig. 9.18 and is a combination of analyses and tests.

9.3.1 Analytical Verification of Space Structures

In order to evaluate a potential mismatch between assumed design loads and actual flight loads on a spacecraft a

Fig. 9.18 A typical structural design validation cycle



coupled launcher-spacecraft dynamic analysis can be performed during the development of a spacecraft. In the case of severe coupling between launcher and spacecraft dynamic modes the actual flight loads could exceed the assumed design loads. This coupled dynamic analysis is generally referred to as a coupled loads analysis and is normally conducted by the launcher authorities. In Fig. 9.19 the coupled launcher-payload system of the Vega launcher and IXV are depicted. Typical output of the coupled loads analysis are the QSL at the center of mass of the spacecraft, the interface forces, the shock response spectrum, SRS, and the equivalent sine response at the interface, and usually some interior responses at critical elements or locations of the spacecraft structure. The equivalent sine can be used to assess the proposed primary and secondary notches necessary in order to not exceed the design loads and local loads respectively.

Besides the coupled loads analysis, the typical structural analyses performed as part of the analytical verification of the structures are

- Modal analysis (to verify frequency requirements).
- Static stress analysis, including thermo-elastic analysis (to derive MOSs).
- Stability buckling analysis (to derive MOSs).
- Transient analysis (to simulate time-domain loads events).
- Frequency response analysis (to validate mathematical models and to simulate sine tests).
- Fatigue and crack growth analysis (to verify the life of safety critical structural elements).

- Acoustic analysis (to check spacecraft response and to derive random spectra).
- Random response analysis (to predict response to random environment).
- Micro-vibration analysis (to predict the effect of reaction wheels, coolers disturbers on spacecraft functional targets like pointing, etc.).
- Dynamic displacement (to verify spacecraft to launcher fairing stay out zone violations).

The un-deformed mathematical model of the Herschel spacecraft used to perform coupled loads and acoustic analysis is shown in Fig. 9.20, along with the same model during a vibration simulation, alongside an image of the Herschel spacecraft on the HYDRA multi-axis vibration table at ESTEC.

The flow chart for finite element analysis is shown in Fig. 9.21.

Structural verification is implemented by following a detailed plan that also includes fracture control activities. For crewed structures, fracture control is applied to all safety critical structures that may pose a safety hazard. For robotic spacecraft structures, fracture control is mainly applied to critical interfaces (e.g. an optical bench attached to the spacecraft platform by means of iso-static mounts) and to pressurized tanks. When required, the specific fracture control activities are specified in a separate plan, normally called the fracture control plan. The verification plans must specify the complete list of activities, analyses and tests to be performed to achieve the certification for flight. In addition, the plans must

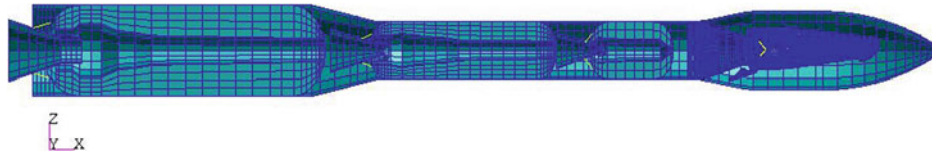


Fig. 9.19 Vega launcher and IXV payload

Fig. 9.20 Herschel spacecraft structural mathematical model (left) and a vibration simulation model alongside an image of Herschel on the HYDRA facility at ESTEC. Image ESA

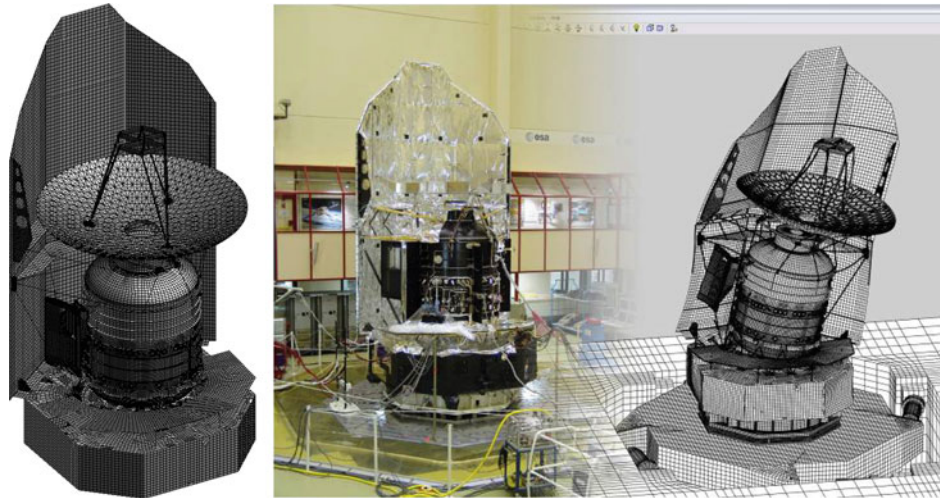
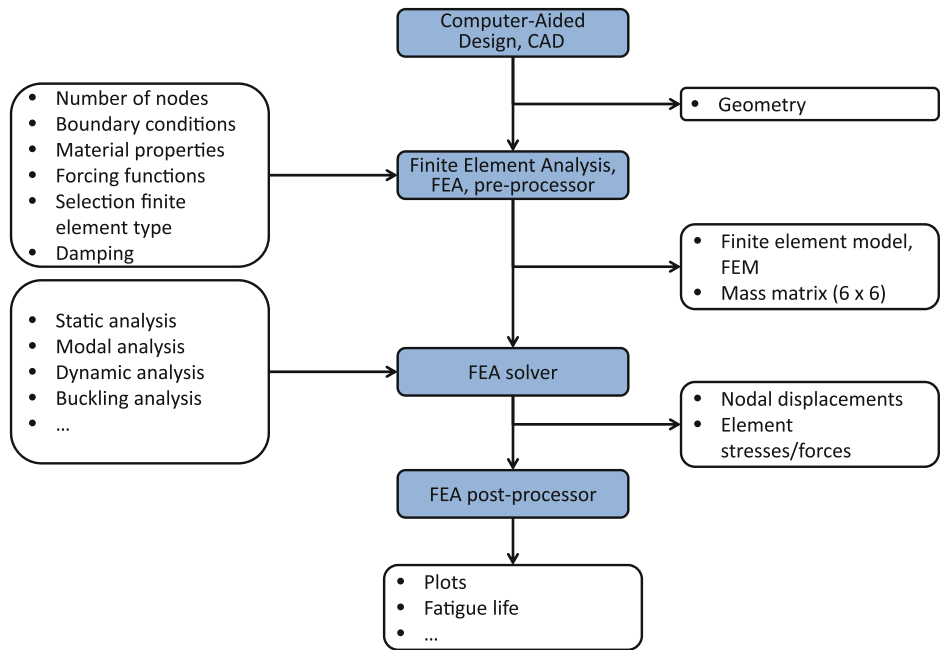


Fig. 9.21 FEM analysis flow chart



specify the requirements applicable to procurement, processes, traceability, and the inspection of materials and structural parts. For hardware that is sensitive to damage during handling, such as glass, ceramics, or composite materials, a damage protection plan must be established to ensure that the critical parts are adequately protected and

handled during the complete mission, including ground activities. Structural verification can be divided in two main activities

- Demonstrate the capability of the design to perform its functions (qualification).
- Certify the quality of the flight hardware (acceptance).

9.3.2 Verification Testing of Space Structures

A very important phase of the verification process is the ground testing of the space vehicle structure. Usually, the testing sequence involves two phases, a qualification phase, aiming at qualifying the system and at obtaining confidence in the analytical predictions (which takes place early enough in the project to limit the risk) and then an acceptance phase, aiming at accepting the end item for the flight. These two types of tests are demanded by the launcher authority in order to show that the structures will survive the launch. In addition, functional tests can be performed to check the performance of complete systems and to verify the unit's specifications.

Qualification testing is done on hardware that is fully representative of the space vehicle structures in terms of design, applied materials, tooling and processes. These tests are often done using structural thermal models (STM), but can also be done on a proto flight model (PFM). The aims of the qualification tests are

- Validation of the design and performance of the space vehicle structure.
- Validation of the compatibility of the spacecraft with the launcher.
- To correlate the finite element model that will be used for the coupled load analyses (CLA) (ECSS-E-HB-32-26A).
- To validate the equipment/subsystem specifications.

Acceptance tests are performed on the flight hardware: flight model (FM) or proto flight model (PFM). The aims of the acceptance tests are

- To verify the FM is free from workmanship defects.
- To confirm that the FM is able to fulfill mission requirements.

To limit the cost of the space vehicle project the proto-flight approach is chosen, in which two testing phases are combined on a PFM that will be tested at qualification levels, but with the duration and the sweep rates at the acceptance level (see Table 9.1). This PFM approach will increase the risk during the project. Special tests can be performed on STM, FM, or PFM models

- To verify the assumptions taken into account in some analyses.
- To verify equipment specifications with complementary analyses.
- To verify the performance of the subsystem.

The following mechanical tests are performed on the structure

- Static load test
- Modal survey test
- Sine vibration test
- Acoustic noise test

- Random vibration test
- Shock test.

9.3.2.1 Static Load Test

The objectives of the static qualification test are

- To verify the structure can withstand the quasi-static loads (QSL) without any permanent deformation (degradation) or failure (strength test). The margins in strength shall be established.
- To gather relevant information about the stiffness of the space vehicle structure for the correlation the stiffness matrix with the structural finite element model.

The quasi-static loads are the dimensioning loads for the primary structure (main load path), whereas the vibration sine loads are important for the secondary structure. The static loads test will qualify the primary structure, its structural connections or joints, and the interfaces of heavy subsystems, e.g. tanks, antennas, solar arrays, and instruments. Several approaches can be applied to perform the static qualification testing, e.g. (1) a static test set-up using a whiffle tree for the introduction of loads, (2) an electrodynamic shaker for sine dwell, sine burst, and sine impulse tests or (3) a large centrifuge.

- The whiffle tree test consists of a set-up where the structure is surrounded by a very stiff rig for the introduction of loads. A second independent rig is used for measuring displacements. Hydraulic actuators at specific points of the structure introduce loads. Combinations of loads can be applied during the static load test and the displacement responses and strains can be measured. Static load test results can be correlated with the finite element analysis. An update in the stiffness distribution (stiffness matrix) can be made.
- For smaller spacecraft, a shaker quasi-static test can be done. This is performed on a vibration table at low excitation frequency to simulate a quasi-static loading via a sinusoidal excitation; however, it can be done in only one direction at the time. The excitation frequency is sufficiently below the first fundamental mode frequency of the space vehicle to justify the assumption that it behaves like a rigid body and to avoid any dynamic amplification. The shaker static test has the advantage that it can be combined with the dynamic testing campaign of the space vehicle. However, because the loading is uniaxial in the case of large structures it is not always feasible to perform this test at a frequency well below the first mode of the space vehicle. Sine-burst is very popular test method. The enforced sine vibration contains only a few numbers of oscillations, therefore hardly any contribution to fatigue will be made (Fig.9.22).

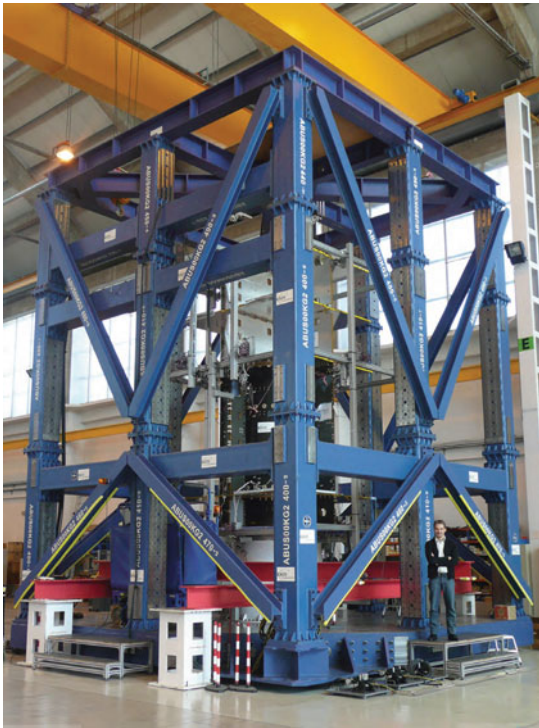


Fig. 9.22 Static qualification test stand (MMST) for the AlphaBus central tube and primary structure at Inta (Spain). *Image ESA*

- A centrifuge test is less common. The mechanical loading is based on centrifugal acceleration. The disadvantage of this method is that it is not possible to stop the centrifuge immediately in the event of failure of the test item. A limited number of measuring channels are possible.

9.3.2.2 Modal Survey Test

The objective of the modal survey test is to validate the dynamic behavior of the space vehicle in terms of natural frequencies, modal damping, mode shapes, and modal effective mass (ECSS-E-ST-32-11C). The results of modal survey test can be used to correlate the mass and stiffness matrix of structural finite element model and to demonstrate compatibility with the launcher minimum natural frequencies requirement. The measured modal damping ratios can be applied for dynamic response analyses in the frequency domain. Attention must be paid to the instrumentation, which should be placed in such a way that all relevant modes can be adequately observed during the test.

9.3.2.3 Sine Vibration Test

The objectives of the sine vibration test are

- To qualify the secondary structures with respect to the launcher dynamic environment.
- To verify the compatibility of the space vehicle with the launcher in terms of frequency.
- To assess damping characteristics (dynamic amplification).

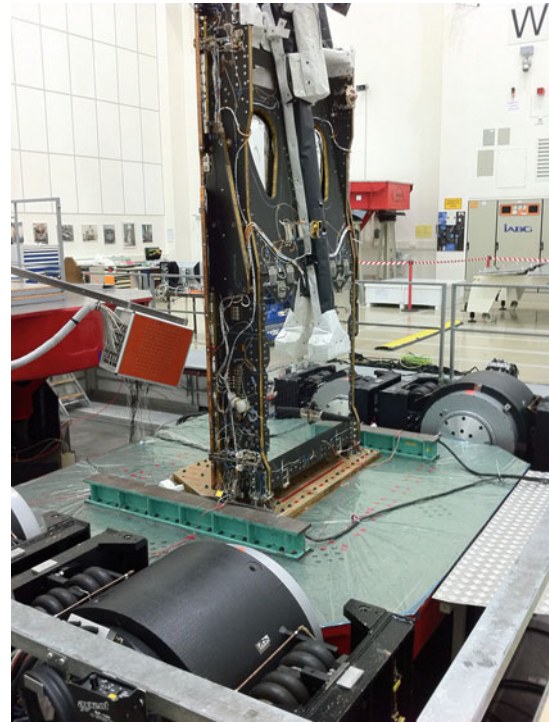


Fig. 9.23 Sine vibration test of the Swarm spacecraft

- To validate antennas, solar arrays, instruments, equipment, etc.

The space vehicle is excited on the shaker table by the specified enforced sinusoidal accelerations. The sine vibration test is usually followed by a functional test. To prevent over-testing, notching of the sine spectrum is often applied in agreement with the launcher authority. An example of a typical spacecraft mounted to a multi-shaker slip table is shown in Fig. 9.23.

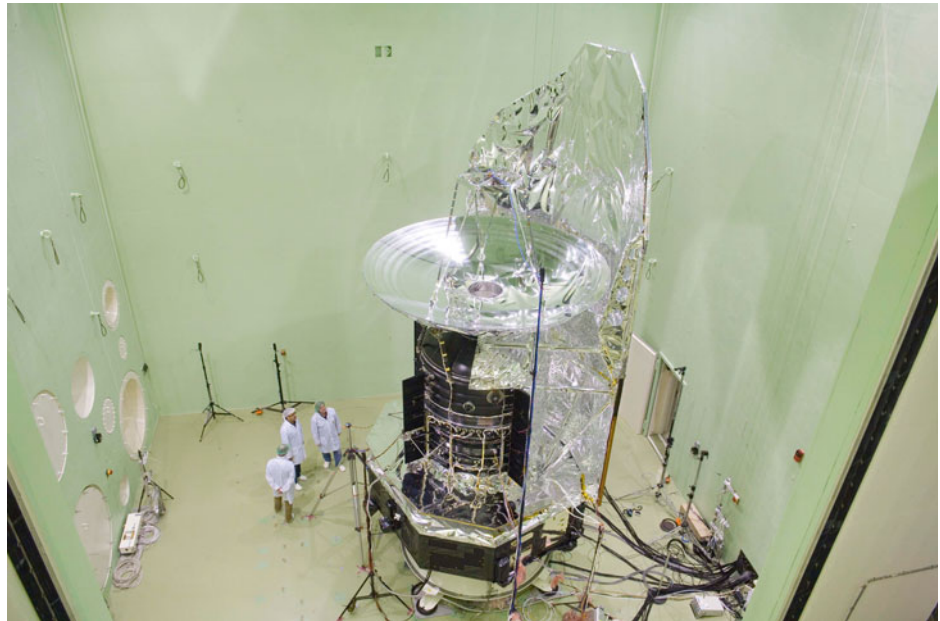
9.3.2.4 Acoustic Test/Random Test

Large areal and lightweight structures (antennas, solar arrays, and radiators) are very sensitive to sound pressure levels (SPL) in the early phase of the launch (i.e. at liftoff). Acoustic noise tests are performed at space vehicle and subsystem level

- To verify the structural integrity against acoustic loads.
- To establish the random vibration specifications (structure-borne random vibrations).

The test is performed in a so-called reverberant acoustic room, as in Fig. 9.24. A reverberant (diffuse) sound field is created inside the chamber by the use of different horns. This sound loading will excite the structure and induce related random mechanical vibrations that are dimensioning loads for equipment (subsystems). Thus, random vibration tests are performed to validate the integrity of equipment. A dynamic shaker will introduce random vibration levels.

Fig. 9.24 The Herschel spacecraft being prepared for tests in the acoustic chamber at ESA/ESTEC. *Image ESA*



Notching is sometimes needed to prevent over testing. In general, random vibrations specifications are covered by the acoustic noise test.

9.3.2.5 Shock Test

Shock loads must be validated by test. The high frequency loads may cause failures in electronic components, mechanisms, valves, etc. Shock loads are mainly introduced by the launch vehicle; e.g. jettisoning the fairing, stage separation, the separation of the space vehicle from the launch vehicle, and the release (deployment) of appendages. Shock tests are thus separated in two categories

- Internal shock test, which is induced by the releasing of different appendages of the space vehicle.
- Launcher shock test, which is induced either by a space vehicle release test (i.e. the clamp band release test) and/or on a dedicated test simulating the launcher shock environment (e.g. the Shogun test to simulate the fairing separation for Ariane 5, the ASAP-S shock kit to simulate the pyro-release shock of the Dassault separation system for micro-satellites on Soyuz, the VESTA device to simulate the release of the Vega fairing).

The shock test will determine the shock transfer functions inside the space vehicle and, by further analyses, the qualification levels for the internal subsystems/equipment can be determined. Space vehicle shock tests are performed on a STM mainly to characterize the transfer functions, and equipment shock tests are performed on EQMs in order not to expose flight hardware to shock tests. Qualification shock tests are always followed by a functional test to ensure the correct performance of the unit. On a case-by-case basis, some other characterization tests can be performed on the space vehicle.

9.3.2.6 Micro-Vibration Testing

Micro-vibration testing is performed to characterize the effect of micro-vibrations induced by moving parts in the space vehicle, such as reaction wheels or compressors (coolers), which can influence the performance of the various instruments on board. This can be done either with a mini-shaker with the objective of retrieving the transfer functions between the excitation point to the instrument (on a STM) in order to verify the analyses carried out, or by activating the different exciters in a space vehicle (i.e. flight model) and observing the disturbances on the instrument directly. The spacecraft is suspended (almost free-free) during the performance of the micro-vibration tests, see Fig. 9.25.

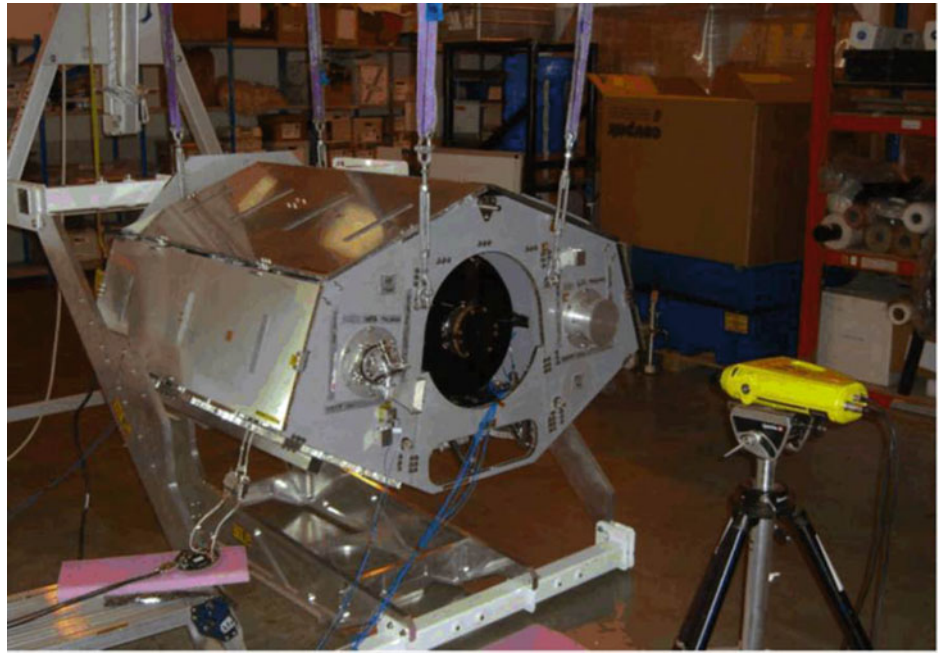
9.3.2.7 Thermo-Elastic Testing

Thermo-elastic test are performed to measure the thermo-elastic deformation caused by temperature gradients, in order to check the effect of deformations on the performance (line of sight) of the instruments. It is not always possible to find a test set-up representative of the real thermal environment, so this test is mainly used to verify the FEM used to analyze thermo-elastic effects.

9.3.2.8 Crewed Space Flight Structures

For crewed space flight, qualification is performed to load levels that are higher than the maximum expected flight levels. This has to be performed both at component and system levels. Qualification of a crewed structure must be performed by a series of tests and analyses. The following list shows the typical tests that are performed for qualification purposes

Fig. 9.25 NigeriaSat-2 structural engineering model undergoing micro-vibration testing. *Image SSTL*



- Static strength test
- Random vibration test
- Shock test
- Acoustic pressure test (capability to sustain acoustic load)
- Acoustics emission test (for human space flight, to verify that the life support hardware does not exceeds the audible noise requirements)
- Microgravity (micro-vibration) test
- Burst pressure tests.

Of course, the list of structural tests can be tailored for the specific application, depending on the structure (pressurized human modules, glasses of a camera, pressurized systems, etc.). It is not always possible to reproduce on the ground all the loading conditions that will happen during the mission, so analyses based on test-correlated models are applied to complement a program of verification by test. Verification by ‘analysis only’ of crewed structures is also possible. In this case, the analysis methods are validated by dedicated tests and applied within their validated domain. Acceptance of a flight model is performed by a series of tests intended to screen for possible defects and bad workmanship, and also to characterize the as-built capabilities

- Random vibration test
- Acoustic pressure test
- Acoustics emission test (for human space flight, to verify that the life support hardware does not exceeds the audible noise requirements)
- Microgravity (micro-vibration) test

- Proof pressure tests
- Inspections (e.g. dye penetrant inspections on finished metallic parts).

Acceptance tests are performed at the maximum expected flight loads and are to check for workmanship defects. Maximum flight expected loads can be exceeded on a flight model only when the test is to screen for the absence of defects that would cause the failure of the structure under nominal loads.

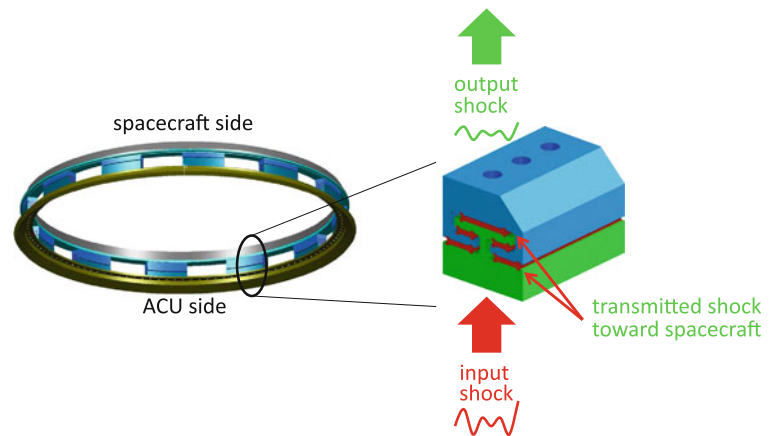
The preferred option for structural verification is to perform qualification tests on prototypes: flight-like models that will not be used for a mission. Such models are referred to as qualification models. When this is not possible, some qualification tests can be performed on the flight model. This increases the risk of damage to the vehicle that it is tested at qualification levels, without actually confirming the full status of the flight hardware; therefore it is not a commended option. It is normally only used when sufficient experience with similar structures exists. The structure has to be designed to withstand qualification loads.

Safety factors to be applied in the verification of crewed structures are usually higher than those for robotic missions. Common ultimate safety factors for metallic structures, or composite structures in no-discontinuity areas, are 1.4 for launch loads and 1.5 for in-orbit loads. Such safety factors are typically increased to 2.0 for discontinuity areas in composite elements. The minimum safety factor for shatterable materials is 3.0. Ultimate safety factors for pressurized structures

Fig. 9.26 Dampers used in space structures. *Image Astrium*



Fig. 9.27 Shock attenuation system for spacecraft and adapter. *Image Astrium*



can go from 2.0 for habitable modules up to 4.0 for flexible lines in fluid systems. In the case of analysis-only verification, the safety factors have to be increased.

9.4 Loads Attenuation: Damping and Isolation in Space Vehicle Structures

Damping measures the dissipation of energy that occurs during vibration or shock of a structure and is a structural characteristic that limits the magnitude and duration of response to input forces. Vibration isolation aims to reduce the effects of vibration on a given structure by isolating it from the source of vibrations. The function of an isolator is either to reduce the magnitude of motion transmitted from a vibrating support to the equipment or to reduce the magnitude of force transmitted from the equipment to its supporting structure. The performance of an isolator to steady-state sinusoidal vibration may be evaluated by the absolute transmissibility (ratio of vibration amplitude of the equipment to the vibration amplitude of the supporting structure), the relative transmissibility (ratio of the relative deflection amplitude of the isolator to the displacement amplitude imposed at the supporting structure), or the motion response (ratio of the displacement amplitude of the equipment to the quotient obtained by dividing the excitation force amplitude by the static stiffness of the isolator).

Sources of vibration can be either external (e.g. launcher) or internal to the spacecraft (moving parts, mechanisms, and so forth). In the latter case, one usually describes the vibration as disturbance vibrations. Those vibrations can be detrimental to the spacecraft integrity due to their amplitude (e.g. the launcher dynamic environment or separation shocks) or to a sensitive receiver performance (e.g. an optical payload sensitive to micro-vibrations). Attenuation of those loads is therefore a constant preoccupation in the design of spacecraft structures.

Generally speaking, depending upon the design principle, vibration control systems can be classified as either damping or isolation. Each class can be further categorized as an active or a passive control system. For passive damping systems, the idea is to dissipate the vibration energy at various locations distributed on the structure in order to control the vibration amplitude of any resonance that may lead to excessive stresses (see Fig. 9.26). Damping treatments require the selection of appropriate materials, locations of the treatment, and choice of configurations, to assure the transfer of deformations from the structure to the damping elements. However, the necessary extra mass constrains the use of passive system over large areas. Moreover, passive damping limits the amplification at resonance but tends also to reduce the high-frequency attenuation, and this could turn to be problematic with harmonic sources like reaction wheel assemblies (RWA) and/or cryocoolers.

Passive vibration isolation is implemented by placing the equipment on appropriate mountings (see Fig. 9.27). A variety of implementation devices can be used, including viscous-elastic materials, springs, soft materials, hydraulic dampers and pneumatic isolators. For the Hubble Space Telescope, a passive isolation system designed by Honeywell that employed a viscous fluid-damped isolator was used to mitigate the effects of the RWA. Another example of an isolator used in several space missions is the Honeywell D-strut, which uses a bellows system with a viscous damped-fluid similar to the one used on the Hubble Space Telescope. Rather than implementing sophisticated devices or materials, secondary structures can be properly designed to mitigate or reduce mechanism-generated disturbance: modifying equipment locations in order to increase the decoupling between the source and the receiver only by modifying the load path; modifying the location of a particular stiffener; or designing mechanical junctions between plates and beams, etc., to reduce the coupling between the substructures.

Active systems usually supply additional power to the system in order to modify its own dynamic behavior. Despite the complexity of active systems, in some cases they represent the only solution to achieve the requested levels and bandwidth of attenuation. They are usually based on local feedback between a co-located sensor and actuator, and they generate an output force proportional to the measured function. One of the main control issues is related to the finite bandwidth of controls and ‘spill-over’ effects. The principle of another possible approach, active compensation through centralized adaptive control, is to generate vibrations using actuators in order to destructively interfere with the disturbances to cancel the jitter of the sensitive instruments. Active isolation is able to simultaneously achieve a low amplification at resonance and a large attenuation at high frequency. In [13], an active six degrees of freedom local vibration isolation applied to a flexible space telescope has been studied for attenuating high-frequency disturbances. A six-axis single-stage active vibration isolator based on a Stewart platform has been developed for space application in [14]. And, for spacecraft whose primary structure are struts, actively controlled struts can be used to attenuate loads and/or disturbances.

Active damping has been implemented in the piezoelectric Stewart platform. This uses a stiff active damping interface as a support for payloads and consists of a Stewart platform with six degrees of freedom, where each leg of the active interface is made of a linear piezoelectric actuator. A mixed control approach may be needed in some cases. For example, an integrated active damping device (IADD a piezo based device) arranged in hexapod configuration has been studied for spacecraft requiring high pointing

accuracy. Similar approaches, where a frame platform has been designed to act as a supporting structure for devices like RWAs, and where a vibration isolation interface between the spacecraft and terminal with appropriate transmissibility characteristics is designed using a monolithic passive flexible element (MEDI), can be used to control stiffness and damping. Passive springs and viscous damping from a linear motor, which may provide activator and sensor functions, provide stiffness.

9.5 Space Vehicle Materials and Processes

High quality materials and processes are required for assuring the performance and reliability of aerospace components, and have contributed to the success of many space missions. The requirements for materials and processes to meet the levels of quality assurance and control imposed by aerospace industry and agencies are defined in specific standards, e.g. ECSS-Q-ST-70C and NASA-STD- (I)-6016.

9.5.1 Selection

Selection of materials for space application requires knowledge of the relevant engineering properties and environments to be endured during the mission lifetime. In assessing the competence of a candidate material for a specific use it is necessary to consider the associated process, targeted application, and respective environment. Therefore, properties such as specific strength and stiffness, fracture toughness, fatigue resistance, stress corrosion resistance, thermal resistance, sublimation, and erosion may be relevant. Other selection criteria that may also be considered are ease of manufacturing and ability to be repaired.

Selected and applied materials in flight hardware must resist ground, launch, and on-orbit environments. Characterization of the performance in anticipated environments is therefore essential. Recommended further reading can be found in [15–17].

9.5.2 Environmental Effects

The evolution of material properties in space is of primary interest in the design of space vehicle structures. Understanding the space environment and the influence on the materials of degradation or even loss of performance throughout the service lifetime, is of great importance. In this context, countermeasures for thermal radiation, vacuum conditions, micrometeoroids, and space debris are important design parameters.

The expected temperature range of the structure is normally also an important factor. Looking to the extreme temperatures, the materials selection will yield different results for cryogenic tanks on the one hand and on the supporting structure of the thermal protection system of reentry vehicles on the other. Temperature variations within the structure may cause unwanted thermal distortions. Many thermal cycles from low temperature to high temperature are of importance in the verification process. This is also important in composite materials, due to the different thermal responses of the fibers and the matrix.

The occurrence of atomic oxygen, corrosive environments, fluid compatibility issues, vacuum outgassing, moisture effects (absorption/desorption) etc., occurring under specific circumstances, all require proper attention. Selecting the most suitable material for a certain application is not always straightforward. In many cases, it is even possible to have more than one solution.

9.5.3 Metallic Alloys: General

Metallic alloys have been the primary choice since the early days of space exploration. A variety of aluminum alloys, ranging from 7020 for welded structures to 2000 (2024) and 6000 (6061/6063) series plus 7075 for un-welded applications, have been used predominantly. Aluminum alloys are often used in plate, shell structures, truss elements, face sheets, and the core of sandwich structures. For structural parts that are subjected to very demanding thermal environments (e.g. nozzle parts, etc.) conventional titanium alloys were the natural choice. High-strength titanium alloys are applied in heavy load-carrying structures such as attachment fittings, fasteners, and pressure vessels. High-strength steels are used in support structures and solid rocket motor cases.

9.5.4 Advanced Metallic Alloys

Over the years, the stringent requirements imposed by the ever more competitive aerospace industry and the increased demand for lighter and stiffer structures have stimulated technology and research parties to seek advanced material solutions. The effects have been visible in the paths followed by industry in terms of research, which came up with innovative solutions, and even in terms of manufacturing strategies, where low-cost manufacturing and unitized (cheaper to assembly) parts are gaining more importance. Materials with enhanced mechanical behavior, optimized performance, lower density and, at the same time more cost-effective, have emerged.

Some development trends are easily identified. Research on aluminum alloys is progressing on different levels, ranging from improved-strength properties (e.g. zinc aluminum alloys, Al-Zn, and alloys of aluminum and lithium, Al-Li) to damage tolerance improvements (e.g. aluminum-copper, Al-Cu, and Al-Li alloys) or even the development of high-temperature aluminum alloys. In reality, a new generation of Al-Li alloys are conquering the aerospace market. These low-density alloys are attractive to the aerospace industry due to their substantial reduction of mass, improved stiffness, and good welding properties. In addition, these alloys present very good resistance against fatigue crack growth, and are therefore suitable for critical components that demand good damage tolerance. The application of Al-Li alloys is very profitable because production costs are low compared to the high investments to set up a fiber-reinforced composites production line. For further information following references are of interest [18, 19].

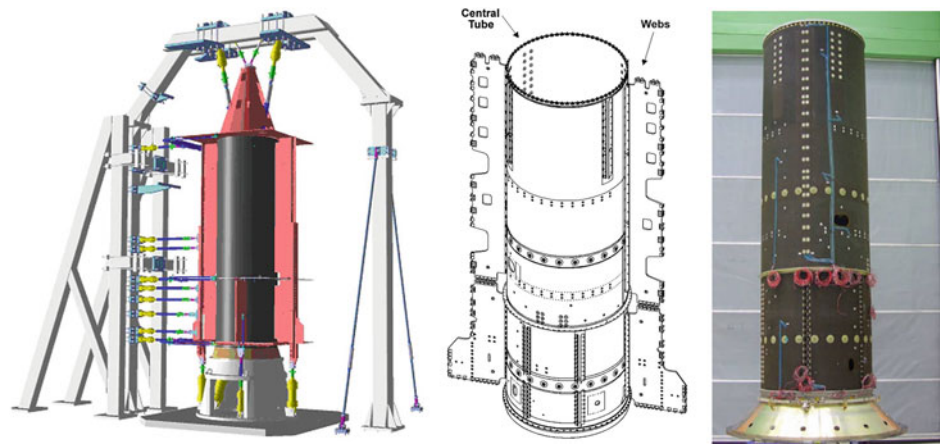
9.5.5 Composites

A composite material is composed of at least two elements, with the fibers and the matrix working together. The mechanical properties of engineering materials usually depend on the number of defects within the structure of the material, and are much lower than theory would predict. It has been found that fine particles and fibers have properties much closer to their theoretical maximum. For instance, the tensile strength of pure silica glass is about 50 MPa, but in the form of a fiber the tensile strength could be in excess of 1,400 MPa. In order to utilize these properties the applied load should be transferred, via its matrix, to the embedded fiber reinforcement

The major composite classes include polymer matrix composites (PMC), metal-matrix composites (MMC) and ceramic matrix composites (CMC). Further, composites can be classified according to the reinforcement form—particles, platelets, whisker or short fibers, or continuous fiber (uni-directional, UD, laminated or woven composites; including braided, knitted and tri-axial architectures).

- PMCs consist of polymer-based resin (thermoset or thermoplastic) as the matrix, and a variety of fibers such as glass, carbon, and aramid as the reinforcement.
- MMCs consist of a metal such as aluminum as the matrix and are reinforced with fibers or particles that can resist the manufacturing process such as silicon carbide or high melting-point metal.
- CMCs are mainly used in very demanding applications (high-temperature environments, high stability). The most common are carbon and/or silicon carbide matrices reinforced with carbon fibers. Other applications require composites systems such as a ceramic as the matrix

Fig. 9.28 Mecabus/Spacebus 4000 Central Tube. Static load test rig (left); schematic of central tube with north and south webs (center), image of central tube (right). Image RUAG ThalesAlemia (left and center) and ESA (right)



reinforced with short fibers or whiskers such as those made from silicon carbide and boron nitride.

The most widely used composite materials for space applications are PMCs with continuous fiber reinforcement and an epoxy or cyanate ester matrix (thermosetting resins). Within each of the groups defined above, a large variety of fiber and resin systems are available, for example many different glass fiber and carbon fiber products exist, each with specific characteristics and properties.

Composite materials are extensively used in the space industry, mainly in the form of carbon fiber reinforced plastics/polymers (CFRP). Carbon fibers either as single fiber bundles, known as tows, or as woven mats are impregnated with organic resins to form strong, stiff, and stable structures. These may be used as structural elements, solar panels, antennas and mirror supporting structures. The orientation of the fibers within the material structure allows exceptional control of the properties, and modern composites are approaching the theoretical strength limits for the material combinations.

Typically, two types of carbon fibers exist, PAN or PITCH, depending on their precursor materials. PAN fibers come in two forms: high strength and high modulus. Both forms are based on polyacrylonitrile (hence PAN) but the details of the graphitization processes vary to give the required properties. PITCH-based fibers use spun petroleum pitch fibers as a precursor. The fibers produced are not as strong as PAN-based fibers but are exceptionally stiff; a modulus of almost 1,000 GPa can be achieved in comparison with almost 600 GPa for the best commercially available PAN-based fibers. The PITCH-based fibers are brittle and can break easily when bent, making handling and use very difficult. In addition to high stiffness the PITCH-based fibers have a very high thermal conductivity, which has resulted in their use in thermal management systems where individual broken fibers are less of an issue. Typically, the space industry uses PAN-based fibers with either high modulus for spacecraft or high strength for launchers.

Composites are normally preferred in mass and stiffness driven applications, such as primary structures (see Fig. 9.28) and payload adapters. Most composites are formed using a thermosetting resin in combination with high stiffness carbon fibers. However, the use of thermoplastic resins is emerging, and will likely become the reference method for the matrix element within carbon fiber reinforced structures. Thermoplastic resins show better mechanical properties than thermosetting ones, although the manufacturing of dimensionally accurate structural elements using this type of resin remains a challenge. The excellent properties and potential mass savings make composites a very attractive solution for many applications. Spacecraft primary structures like platform panels, central tubes, and secondary structures like payload panels, sunshields, antenna reflectors, and solar array substrates are made of composite materials.

The range of applications for composites is extending, and nowadays they are also applied to pressure tanks, either in COPV form (composite overwrapped pressure vessels) or as composite tanks (e.g. for cryogenic applications). Several developments are being investigated in these areas, to identify possible concepts (e.g. simple skin, sandwich construction, multiwall) and overcome technical difficulties such as compatibility issues, damage tolerance and health monitoring, thermal protection integration, reusability, etc.

Composite materials offer new possibilities to associate function, complex forms, and materials, and to better satisfy customer requirements (weight, functional, etc.) where the application of metallic materials is difficult. Composite materials have many functional advantages: lightness, mechanical strength, reduced maintenance, and complex forms. Thanks to their mechanical properties they allow an increased lifetime for structures, and show a greater impact resistance. In addition, fiber composites allow considerable weight reductions of 10–20 % compared to classical metallic materials due to the exploitation of anisotropy.

In the field of aeronautics and space vehicle systems, materials used for structural applications are typically carbon fibers and thermoset resin. The future use of carbon fibers and thermoplastic resins will further enhance the advantage of composite materials for space structures as compared to metallic ones, without leading to a prohibitive cost.

9.5.5.1 Ceramics and Glass

Ceramic materials show very interesting mechanical properties, such as high stiffness-to-mass ratio, good stability (due to the low coefficient of thermal expansion) and often high-temperature resistance. However, their brittleness and unforgiving behavior cannot be disregarded, and this drives most of the mechanical design and verification process. The strength of ceramic materials is very much dependent on the surface condition and the distribution of (strength limiting) flaws. For this reason, there is a size effect to be accounted for as ceramics exhibit sensitivity to the volume under (high) stresses. In addition, some ceramics are sensitive to 'slow crack growth' or 'static fatigue', which is a strength degradation phenomenon that occurs under sustained loading and in the presence of an aggressive environment (e.g. humidity). Good stability and a high stiffness-to-mass ratio makes ceramics an almost natural choice for optical structures such as space telescopes, optical benches, mirrors, and scientific instruments.

SiC 100 is a ceramic material produced by bonding silicon carbide (SiC) grains by sintering, and has been used in a number of space missions. SiC has a rather low coefficient of thermal expansion and shows high stiffness down to cryogenic temperatures, which is ideal for large telescopes or stable optical benches at low operating temperatures. Structural components (optical bench, struts) can also be manufactured using SiC. Note that SiC has isotropic properties. Large temperature variations in a telescope made completely of SiC will not affect the optical performance. Additionally, SiC is not susceptible to static fatigue. It has been used to build the telescopes of Herschel and Gaia (Fig. 9.29), both of which were state-of-the-art of large light weight, highly stable structures.

CeSiC[®] is a ceramic material that incorporates fiber reinforcements. The type of fiber (specific fiber placement or chopped short fibers) will introduce non-isotropic characteristics. Flight representative demonstration models of relatively large dimensions have already being manufactured. HB-CeSiC is made up of carbon fibre with silicon infiltration, and is used for mirror applications.

An example of a more glass-like material is Zerodur. Most companies manufacturing space and ground-based telescopes have extensive experience of manufacturing and polishing mirrors in Zerodur. Young's modulus and the strength of Zerodur are relatively modest, so Zerodur

mirrors are generally supported by a lightweight back structure to reach a compromise between mass and stiffness. Using an ultrasound milling machine and optimized mechanical design, specific weights as low as 40 kg/m² can be achieved. Lightweight Zerodur mirrors are still competitive with SiC for telescopes demanding image quality with diameters up to 2-m, due to its lower cost, shorter manufacturing schedule, and risks. The mass of a 2-m Zerodur mirror is approximately 150 kg, while offering both good mechanical behavior and satisfactory image quality. However, glass-like materials such Zerodur show some limitations

- Low thermal conductivity and low Young's modulus limit the potential mass-saving applications (the thermal control axial gradient detrimental is for Earth observation).
- Time consuming manufacturing.
- Joining of parts/segments.
- The applicability of a honeycomb concept to large concave (or convex) mirrors still requires significant developments.

Moreover, the proven performance of some ceramics under extreme temperatures, with some ceramics qualified to temperatures well above 1,200 °C, makes this material very suitable for thermal protection systems of reentry vehicles.

9.6 Manufacturing and Assembly of Space Vehicle Structures

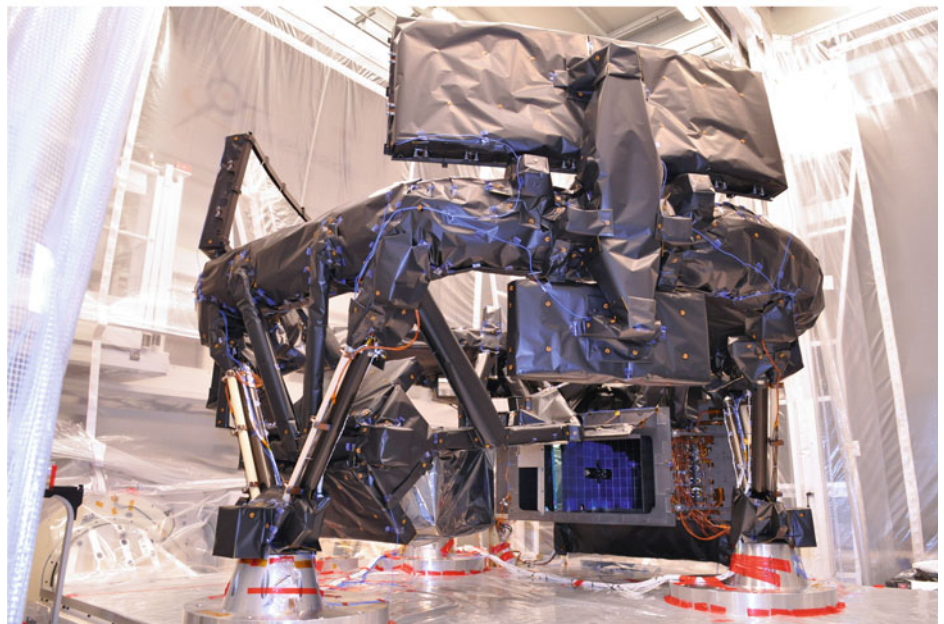
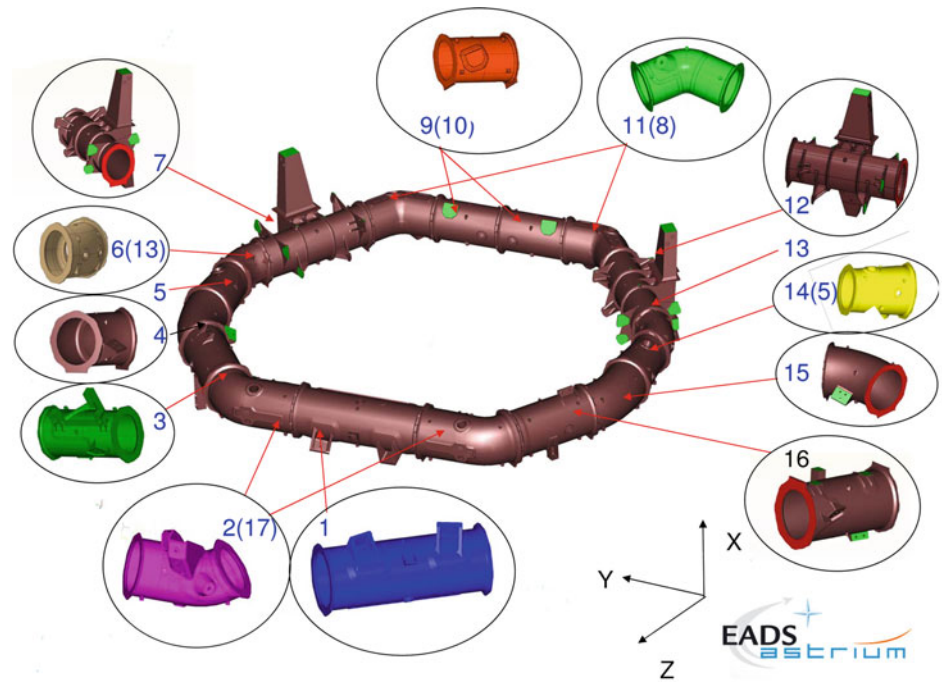
Manufacturing of structural parts varies depending on the type of structure and the material used to build it.

9.6.1 Metals

Machining is the most frequently applied manufacturing method for metals. The process consists of removing material by means of cutting, milling, drilling, or grinding tools. Most machining operations are automated and are supported by CAD/CAM tools.

Forming is one of the most economical methods of fabrication. The most limiting aspect of designing formed parts is the bend radius, which must be large in order to limit the amount of plastic strain in the material. Superplastic forming and diffusion bonding, at temperatures up to 1,000 °C, can produce complex components but only for titanium alloys since others are prone to surface oxidization that inhibits the diffusion bonding. Spin forming has been successfully applied for manufacturing aluminum pressure vessels.

Fig. 9.29 The GAIA payload module. The 3-m diameter, quasi-octagonal torus supports the two telescopes and the focal plane assembly, and is composed of 17 individual custom-built silicon carbide segments (*top*); and the fully integrated GAIA proto-flight payload module (PLM) undergoing acceptance vibration testing on an electrodynamic shaker at the facilities of Interspace in Toulouse, France (*bottom*). *Image* EADS Astrium



Forging, in which structural shapes are created by pressure, is well suited for massive parts such as load-introduction elements. Where large sheets have to be manufactured (as for the skin of main thrust cylinders) chemical milling is frequently used in order to reduce the thickness of the sheet where possible. This process is more reliable than machining when processing very thin elements, but adequate tolerances have to allow for various inaccuracies as thickness variations of the original piece of material are achieved by masking.

Casting can be used to produce parts of complex shapes. However, the quality of a casting is difficult to control because gas bubbles can form as the material solidifies, resulting in porosity. Material strength and ductility are not as high as with most other processes. Also, the application of a high pressure to a metal structure, taking the internal parts to yield and resulting in internal compressive residual stresses after pressure relaxation, is a manufacturing method known as autofrettage and is used to manufacture pressure vessels and other pressurized elements.

Recently the implementation of additive manufacturing (AM) has opened a new possibility for the manufacturing of complex structural parts, either metallic or non-metallic.

9.6.2 Composites

The processing of composites starts with raw materials rather than semi-finished products as for metals, hence it is possible to produce fairly complex parts as a single piece (unitization) and achieve savings on manufacturing and assembly costs. Impregnated tapes are the most widely used precursor for composite manufacturing. Fibers can be woven to form fabrics. The key parameters when specifying the material are the types of fibers and resin, content, tack, and drape.

The fiber will control the major mechanical properties of the part, such as its strength and stiffness. The resin, which binds the fibers together, will determine the maximum temperature under which the part can be safely used. It will also influence moisture absorption and desorption, with possible effects on geometrical distortions under thermal and moisture cycling.

Tack is a measure of how much pre-preg sticks to it and to other layers. Pre-pregs that have too much tack can be difficult to handle because misplaced layers are difficult to reposition without disrupting the resin or fiber direction. Pre-pregs with little tack are difficult to keep in place as more plies are applied. Lack of tack often indicates the pre-preg resin has cured beyond an acceptable limit, with the result that the composite part will not cure properly.

Drape is the ability of the pre-preg to form around contours and complex shapes. The fiber material and the diameter of the filaments, as well as the cross-section of the tow, influence drape. It is also conditioned by the weaving pattern. For flat panels a low tack and drape are acceptable, but for complex shapes and cavities, tack and drape keep the composite in place during laying-up and preparation for curing.

Manual lay-up is a costly technique, but it is well adapted for the production of single parts of very small series. When the number of identical parts to be produced increases, filament winding, resin transfer molding, and braiding, are cost-effective alternatives.

Filament winding consists of wrapping bands of continuous fiber or strands or roving's over a mandrel in a single machine-controlled operation. A number of layers of the same or different patterns are placed on the mandrel. The fibers may be impregnated with the resin before winding (wet winding), pre-impregnated (dry winding), or post-impregnated. The first two winding sequences are analogous to wet or dry lay-up in the reinforced plastic

fabrication methods. Curing the resin binder and removing the mandrel completes the process. Machining or grinding is usually not necessary.

Resin transfer molding is a closed-mold low-pressure process. The fiber reinforcement is placed into a tool cavity, which is then closed. The dry reinforcement and the resin are combined within the mold to form the composite part. This process allows the fabrication of composites ranging in complexity from simple, low-performance small parts to complex elements of large size. The advantages are a very good surface finish and good tolerances. The significant cost of the mold is a drawback that often disqualifies the process for small series production.

In braiding operation, a mandrel is fed through the center of a braiding machine at a uniform rate and the fibers or yarns from the carriers are braided around the mandrel at a controlled angle. The machine operates like a maypole, with the carriers working in pairs to accomplish the over-and-under braiding sequence. Parameters in the braiding operation include strand tension, mandrel feed rate, braider rotational speed, number of strands, width, and the perimeter being braided. Interlaced fibers result in stronger joints. Applications include lightweight ducts for aerospace applications.

Automated tape laying machines have been introduced progressively in order to gain reliability and cost-benefits. Today this technology is mastered for the thermoset resins used in aeronautic and space structures. The Ariane 5 inter-stage structure and payload adapter are some examples. Thermoset materials are used to a large extent, and often components are manufactured using automated tape layer machines and autoclaves. Restrictions arise, however, for very large components due to the autoclave size.

High-performance thermoplastic composite materials have the potential to improve aeronautical structural efficiency, and to reduce manufacturing and in-service costs. Evaluations for space applications have shown the utility of these materials associated with an out-of-autoclave consolidation. Manufacturing of thermoplastic composites using automated tape laying also has a great potential to reduce manufacturing costs, schedule, and risks. Integrating heating on the machine during lay-up and in situ consolidation would be the next step to reduce manufacturing cycles and cost.

Fiber steering is a method of construction for fiber-reinforced composites that allows the unidirectional fibers to be aligned along curvilinear paths. An advanced tow placement machine steers the fibers along the desired paths through computer-controlled trajectories. This allows defining and controlling the stiffness, the density, and the CTE of a panel as a function of the direction and the location of the fibers.

9.6.3 Joining

The assembly of the different structural parts of a space vehicle is achieved by joining them to each other by fastening, riveting, bonding, welding, soldering, or brazing.

Fastening is a frequently used joining method. Torqueing is normally the method used to install a fastener, although there are other methods in which the preload is applied by stretching the fastener without torqueing it. The installation torque is the addition of the seating torque, which produces the desired joint preload, and the running (also called locking) torque implemented to avoid the fastener loosening due to vibrations and suffering fatigue. A fastened joint can be assembled and disassembled a few of times without compromising the preload.

Riveting is mostly used when the joint is designed to carry mainly shear loads. Unlike fastened joints, riveted joints are less suited for disassembly.

Adhesive bonding is used mainly to join composite or ceramics elements. Joining is normally achieved by the use of thin adhesive films. Susceptibility to shock, vibration, and thermal loads must be considered in the design.

Joining by welding is achieved by melting the two parts to be joined and adding a filler material to form a pool of molten material that, after cooling, becomes a rather strong joint. However, when soldering or brazing two parts a lower-melting-point material is used between the parts to be joined without melting them. The residual stresses that are built in during cooling after welding, soldering, or brazing could be relevant in magnitude and must be considered in the verification of the joint or decreased (e.g. by thermal treatment).

9.7 Space Vehicle Mechanisms

Mechanisms are essential for achieving various functions of a spacecraft mission, such as deployment of appendages, high stability pointing and scanning, reaction and momentum wheels, robotics and exploration tools, to name a few. In addition, mechanisms are implemented in launchers and reentry vehicles for thrust control, actuation of control surfaces and landing systems.

In space vehicles, mechanisms are generally not redundant and are therefore considered to be single points of failure. Furthermore, often the operation of a mechanism occurs in a harsh environment, and after a long period of storage or after along flight, so they require a careful and robust design and verification process in order to achieve the necessary reliability.

9.7.1 AOCS Sensors and Actuators

With modern attitude control subsystems, there is a large variety of sensors available to provide the necessary measurement signals such as spacecraft angular position, rate and acceleration. Typical examples are star trackers, Sun & Earth sensors, rate/integrating gyroscopes (gyros), accelerometers, magnetometers, and combined/multi-axis sensor equipment like inertial measurements units (IMU). Furthermore, specialized optical sensors and cameras are also used for navigation purposes.

In the early 1990s, many gyros were still based on mechanical technology, essentially floating gyros and dynamically (or dry) tuned gyros (DTG). However due to reliability problems, most sensors are not based on 'classical' mechanical operating principles anymore, for instance fiber optical gyros (FOG) or hemispherical resonating gyros (HRG). Nevertheless, certain functions in the design of individual sensor types cannot be accomplished without dedicated mechanisms, for instance the scanning mirror assembly of Earth sensors or caging mechanisms for proof mass based accelerometers. Such special-purpose accelerometers may also form a scientific payload, as is the LISA Pathfinder spacecraft, sometimes with sophisticated compensation of magnetic field effects. In addition, it is worth noticing that micro-electro-mechanical systems (MEMS) are utilized for sensors where adequate, e.g. in MEMS gyros.

9.7.1.1 Reaction and Momentum Wheels

In many spacecraft, reaction and momentum wheels are used as actuators for attitude control. Reaction wheels are designed to operate over a wide speed range, including speed reversals. By controlled acceleration or deceleration of the wheel, a reaction torque is applied on the spacecraft platform for the controlled exchange of angular momentum between the satellite and the wheel. By this approach, slew maneuvers and attitude stabilization via the rejection of external disturbance torques can be performed. Momentum wheels are typically operated at a fixed or nearly fixed speed, and are used to provide a momentum bias to a spacecraft to ensure gyroscopic stabilization.

A set of three reaction wheels mounted in an orthogonal configuration can provide attitude control about all three axes of a spacecraft. However, often a skewed configuration of four wheels is used to build in adequate failure tolerance at the system level against the malfunction of any single reaction wheel in the set. In accordance with needs for spacecraft of different types and sizes, there is a large range of wheel products in terms of angular momentum capacity and reaction torque.

The main subassemblies of reaction and momentum wheels comprise an inertia rotor, a bearing unit for rotor suspension, an electric motor with the associated drive, control & interface electronics, a housing with the mechanical and electrical interfaces, and auxiliary parts such as sensors for monitoring purposes. The electric motor to drive the wheel is often implemented as a brushless DC motor. The inertia rotor comprises a carefully balanced mass in some rotationally symmetric shape. In most wheel designs, it is suspended by precision ball bearings.

There have been various development activities on magnetic bearing suspension, and a number of operational spacecraft use magnetic bearing wheels. However, because of the added design complexity, cost, difficulties with ground testing, and other aspects, they have not been commercially attractive up to now and have been limited to missions such as Earth observation spacecraft. Wheel technology development is pursued in response to the requirements for present and future attitude control systems and space missions and includes advanced wheel internal control schemes, drive electronics with increased performance and alternative tele command & telemetry interfaces, and the reduction of wheel-induced micro-disturbances.

9.7.1.2 Control Moment Gyroscopes

Control moment gyroscopes (CMG) have been applied for attitude control on large spacecraft such as space stations and military satellites. There is now also an increasing interest in CMGs for agile spacecraft, e.g. civilian Earth observation missions. They offer the opportunity to reduce payload mass and complexity, in particular when developing a medium-size spacecraft (up to about 1,000 kg) with high maneuverability.

CMGs comprise a flywheel (usually with constant angular momentum), accommodated on a gimbal structure. Single and double gimbal configurations have been used. However, most of the presently developed CMGs are single gimbal configurations, mainly to avoid complexity.

A gimbal actuator rotates the spin axis of the flywheel, which results in a gyroscopic output torque. This torque is proportional to the flywheel momentum multiplied by the gimbal rotation rate. Therefore, CMGs are particularly suited as high-torque actuators that enable high slew rates of a space platform. For instance, using a reaction wheel with 15 Nms installed momentum and rotating it about the gimbal axis at 3 rad/s, a torque of 45 Nm can be achieved. This output torque level is effectively more than 100 times larger than the typical torque capability of the same wheel in a conventional reaction wheel configuration. However, it was necessary to develop sophisticated attitude control algorithms, in particular to avoid singularity configurations with multiple CMGs (see also [Chap. 12](#)).

Technology development for future applications is focused on even more compact CMGs for spacecraft below 500 kg, on the reduction of micro-disturbances (mainly caused by flywheel rotation), and further improvement of subassemblies such as the control, drive, and interface electronics.

9.7.2 Electrical Motors

An electrical motor is the combination of an electromechanical converter and its controller. There are many types of electric motors. The most relevant for space applications are

- Electromagnetic motors
 - Brushed DC motors (including brush equivalent concepts)
 - Brushless DC motors
 - Stepper motors
 - Voice Coil motors
- Non electromagnetic motors
 - Piezo-electric motors and other (often non-magnetic) working principles, including those used in micro-technology.

The commonly named electrical motor is a device capable of providing elementary motion along one axis within a mechanism. It is a complex assembly of many components. Each component has its own particular technology, but the most fundamental is the frameless electrical motor. The electromechanical converter is composed of a motor integrated to a speed reduction device to provide motion along one axis with the required energy and speed. This motion can be rotational or linear. The main components of an integrated motor are typically a frameless electrical motor mounted in its housing, a bearing assembly, and a shaft. A phase commutation device or a position measurement sensor might also be necessary.

The frameless electrical motor comprises a fixed stator and a moving rotor. The rotor may be internal or external. Other variants are also possible, such as an axial gap with a disk shape rotor and stator. The most common motor is nevertheless the configuration with an external stator with windings, combined with a rotor equipped with magnets mounted on the output shaft of the motor. The shaft rotation is guided by the bearing assembly, which is most commonly achieved by a pair of preloaded ball bearings.

Different types of bearing assemblies can be used for space applications, including ball bearings, and magnetic and hydrodynamic suspensions. Each of these requires specific tribological solutions, consistent with space requirements.

One of the major motor cost drivers, and one of the main trade-off parameters for the space mechanism designer, is the motor controller. It provides the motor windings with

the required current in order to generate the motion. The controller can be a simple DC voltage provider, usually including a current limiter, or a complex and expensive electrical power provider, incorporating numerous functionalities. The controller also includes all the software required by the mechanism's movement during the release and operational mission phases. To simplify the architecture of a motor controller, it is possible to consider it as being composed of two different parts: the high power part (the power supply) and the low power part (the signal electronics). The latter covers the commutation, sensor and command signals. The power stage provides the current, and hence the power, to the motor with the appropriate phases (e.g. 2- or 3-phase). The power stage of the controller interfaces with the motor windings (in most cases, with the stator).

The fundamental and specific electrical motor know-how is the capability to design a frameless motor and identify the associated power supply requirements. The remaining motor technologies are encountered in any mechanism, and are therefore not necessarily specific to electrical motors. These include bearings, tribology and electronics design. Due to this complexity, it is a significant challenge for the space mechanism designer to select and procure a technology that will be optimal in terms of motor concepts, performance, reliability, with materials suitable for space applications, and also compatible with the application schedule and overall costs.

Depending on the motor's magnetic concept, the way in which an electrical motor brings about mechanical motion to a mobile payload can be radically different. To simplify the understanding of these differences, it can be stated that some motors are intended to provide forces, while others provide positioning as a function of time. The brushless DC motor is a typical example of the first category, while the stepper motor falls under the second. There are other types of electrical motor technologies, such as the variable reluctance stepper motor, or the induction motor, as well as other varieties that produce linear instead of rotational motion.

Electric motors for space, within motor technology in general, exploit electromagnetic and non-electromagnetic operational principles. Of these two main families of electric motors and actuators, the most important and most conventional are the electromagnetic group of devices. However, the rapid introduction of piezo systems into diverse terrestrial applications means that they must be considered, because some of their characteristics make them attractive for space applications. In this vein, it is important to mention *en passant* magnetostrictive devices, which are similar in characteristics in many ways to piezo systems.

There is a strong trend of development and diversity in piezo electric motors and piezo actuators. Motors ranging

from nano-scale to several cubic centimeters in volume are available in a wide variety of shapes and configurations. The industrial demand for small, lightweight motors in portable technologies will continue to stimulate research and development of new piezo motor designs; although adaptation and qualification will be required before this technology can be used for space applications. Piezo actuators seem ideally suited to sub-miniature mechanisms, which is a niche of space applications. However, piezo motors have much lower efficiency, and shorter lifetime than electromagnetic counterparts. Furthermore, piezo actuators are not seen as a general replacement for small electromagnetic motors, but as solution to specific small-scale actuator applications.

Electrical motors are currently present in a very large number of terrestrial applications: their annual production is in the order of millions of units, while electrical motors flown in space are around several hundred units per year. Therefore, it is clear that the main technology developments in this field come from terrestrial applications. However, some specific product adaptations have been made to fulfill the particular requirements of space applications.

Ongoing improvements of the technology should increase the power/mass and power/volume ratios, improve efficiency, reduce noise and vibrations, increase speed ranges, and increase reliability. Any exceptions to this should come from the developments required for specific applications, such as extreme-temperature motors, motors operating in corrosive environments, etc. Nevertheless, new industrial applications of electrical motors, like the following ones, might stimulate motor improvements in new technology areas that would be beneficial to space applications.

- Environmental and climate concerns are stimulating a resurgence in interest in electric motor design. Particular interest is being shown in designs for solar powered vehicles and electric and hybrid cars. The electric motor is being viewed as a system, in which the controller electronics, the motor efficiency, and the matching of the load to the motor must all be improved and optimized in order to increase system efficiency.
- One of the newest and most vigorous trends is in motor controller electronics for permanent magnet motors, switched reluctance motors, and hybrid systems. This will allow huge flexibility of both speed of operation and frequency of operation.
- Advances in high-purity steels, new permanent magnet materials, soft magnetic materials and conductors and insulation are making possible many new construction geometries and designs.
- Superconducting motors should soon be feasible thanks to new fabrication methods for making high-temperature superconductors (HTS) into flexible wires suitable for winding coils. With transition temperatures now up

to ~ 90 K, cooling is within the range of liquid nitrogen and cryocoolers.

- There is a trend to smaller, high-speed motors, exploiting power = torque \times speed. This is becoming evident in domestic appliances such as vacuum cleaners and cordless drills, which use small motors running at 100,000 rpm, with appropriate gearing.
- Fault-tolerant motor systems are developing well for space applications, especially for launchers, in synergy with the needs of the aviation industry and the military. Mean time between failures of motor systems such as for aircraft flap control are believed to be $\sim 10^5$ h (i.e. about 12 years).
- In this respect there is a trend to replace hydraulic systems of launchers by electrical systems (as has been done for the flap control in the aviation industry with fault tolerant electric motors) in order to save weight and increase reliability.
- There is a trend to develop high-temperature motors. This allows the use of electric motors actually inside jet engines, and in the hot areas of, for example, car engines. Continuous operation at above 350 °C is required for in-engine applications in aviation.
- There is steady progress in low-temperature motors, although issues remain regarding bearings, lubricants, and materials.

9.7.3 Pointing Mechanisms

9.7.3.1 Electric Propulsion Pointing Mechanisms

In order to optimize electric thruster propellant resources, to limit the number of electric thrusters on a platform, and to perform advanced attitude and orbit maneuvers, electric propulsion pointing mechanism (EPPM) are required. EPPMs are used on the majority of the spacecraft using electric thrusters with a power consumption >1 kW. Their main functions are

- To accommodate and support electric thrusters during ground, launch, and on-orbit activities.
- To secure during launch the electric thrusters in a stowed configuration.
- To allow on-orbit multi-axis pointing capabilities of the thrust vector(s) of the operational and redundant (when applicable) thrusters under the control of the attitude and orbit control system (AOCS).
- To provide a dynamic transfer function compatible with the thruster allowable mechanical loads of a thruster.
- To provide a thermal design compatible with thruster and platform thermal requirements during both the operational and non-operational modes of a thruster.
- To accommodate and route the electrical harness and pipes around each rotational axis to ensure adequate margins of life and torque.

- To accommodate electric propulsion system (EPS) ancillary equipment (e.g. a hot interface box, xenon flow control unit).

Generally, EPPMs are composed of the following elements

- A mobile plate supporting one or two thrusters, and thruster interface shims (if any).
- A multi-axis (generally two) pointing assembly including drive units and a kinematics assembly.
- A sensor unit, the complexity of which depends on the AOCS control logic.
- A hold down and release mechanism.
- A flexible thruster supply lines assembly, including supports and protection from radiations a new responsibility of EPPM suppliers.
- A damping system due to the high sensitivity of electrical thrusters to vibrations and shocks.
- Tailored mechanical and thermal interfaces with the thruster(s) and ancillary equipment.
- Supports for EPS equipment.
- Passive and active thermal hardware.

The main technical challenges in the design and development of EPPMs are

- To minimize the mechanical environment at thruster interfaces.
- To operate under a stringent thermal environment.
- To allow the routing and flexibility of thruster supply lines.

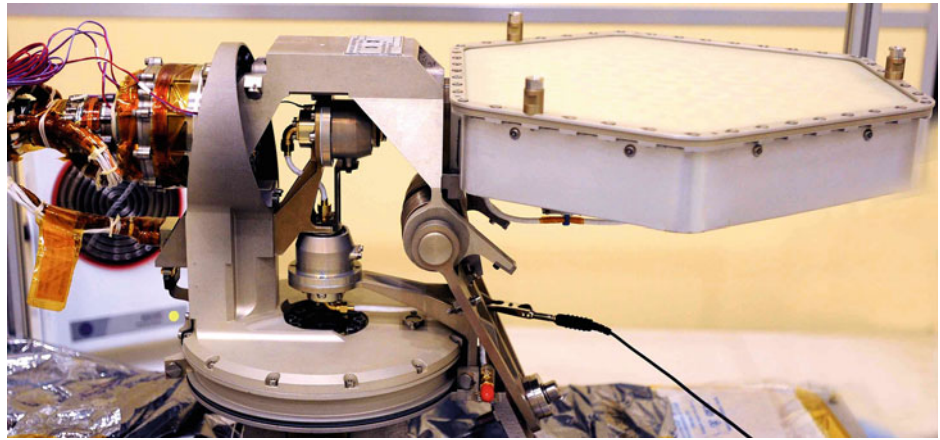
EPPMs must be considered as an enabling technology in order to maximize the performances of electric propulsion systems. EPPMs developments in Europe were initiated in the early 1990s as electric propulsion was identified as the most promising technology to provide the specific impulse needed for future telecommunication, science, and even some crewed missions.

The recent history of electrical propulsion systems has been paved by successes of the Artemis and SMART-1 spacecraft and the Eurostar platform, by the consolidation into reliable system of both the Snecma and Fakel PPS thrusters, and also by the in-orbit degradation observed in Boeing commercial telecom satellites.

The new generation of electric thrusters offer enhanced thrust capabilities that enable more spacecraft maneuvers to be achieved using electric propulsion. This requires extended functional performances of the pointing mechanisms, including complete reorientation of the thrusters in order to achieve various spacecraft propulsion scenarios.

The future objectives are to extend EPPMs and then EPS capabilities in order to achieve momentum wheels damping, east–west station-keeping, orbit top-up, and potentially orbit raising/transfer, orbit inclination changing, and graveyard maneuvers. Future EPPMs will have to accommodate the high-power electric thrusters (typically >5 kW) which are more than twice as heavy, provide improved thermal

Fig. 9.30 High gain antenna, including gimbal pointing mechanism built by SENER for the Mars Science Laboratory. Image EADS Astrium



management capabilities, accommodate stiffer fuel and electric lines, and operate for higher numbers of cycles and cumulated angle. High power and voltage cables are considered as being the driver. However, one of the major enhancements will lie in generating very low dynamic amplification in order to guarantee the thrusters' integrity during launch and shock-generating operations, and in increasing dramatically their pointing range capabilities.

9.7.3.2 Coarse Pointing Assemblies for Optical Terminals

The coarse pointing assembly (CPA) is in essence a two-axis pointing system. It comprises a gimballed mirror driven by independent elevation and azimuth scanning units. Stability is the keyword for the design of the CPA. It must support the mirror without inducing any stresses mechanical or thermal. It must not be sensitive to micro-vibrations originating from the host spacecraft, and it must offer a stable optical platform when rotating the mirror (i.e. a smooth well-damped motion).

The CPA is part of the optical head, the primary function of which is to provide the beam steering and realize a free-space optical link. The other major units within the optical head are the telescope, optical, bench and the fine pointing assembly. The primary role of the CPA within this system is to point the terminal (beam) towards the partner satellite, compensating for low rate disturbances and drift.

9.7.3.3 Antenna Pointing Mechanisms (APMs)

Pointing mechanisms are needed to ensure correct alignment between antenna terminals to guarantee an efficient radio frequency link, either for a space-to-Earth link or for an inter-orbit link. Typically, entire antennas or major elements (like the reflector dish or the antenna radiating horn) need to be oriented in a prescribed direction as requested by the antenna pointing system (APS), and electrically implemented via the antenna pointing electronics (APE). Angular motion normally includes two degrees of freedom, since

rotation around antenna bore sight does not normally need to be actively controlled.

Typical mechanical architectures to implement the two degrees of freedom angular motion are

- elevation/azimuth rotary stages (normally in a elevation over azimuth configuration) capable of large, quasi-hemispherical angular pointing coverage.
- gimballed pivot point with two linear actuators acting as an offset lever arm for smaller angular ranges.

An example of a gimballed antenna pointing mechanism is shown in Fig. 9.30, from NASA's Mars Science Laboratory. It provides communication with Earth on the X-band, without intermediate links.

APMs can constitute an (unnecessary) load-carrying path during launch phases, when the antenna is in its stowed configuration. Consequently, the overall antenna mechanical architecture needs to be carefully conceived to avoid excessive loading at critical locations (like in the APM bearings) due to hyperstaticity. Particularly critical is the interaction with the hold-down and release mechanisms, and to prevent excessive shock loads at separation reaching sensitive APM components. A typical APM configuration includes (for each rotary stage)

- Main bearings to implement the rotational motion
- Gear stage to couple the actuator to the APM output shaft
- Actuator, typically stepper motors typically with integral gear-head
- Angular position sensor
- Twist capsule or cable wraps
- Thermal hardware for temperature control.

In order to transfer the signal/power cables through the APM rotational stages, twist capsules can be used. Since in most cases there is also a need to transfer a radio frequency signal, an APM might include hollow shafts, through which radio frequency rotary joints are implemented. This is particularly true for high-frequency applications (from X-band up to Ka-band and higher) where coaxial cables would not be an option due to high losses. For limited angular motion

(typically when gimballed antennas plus linear actuator are used) the radio frequency signal can be transferred by means of sections of flexible wave-guides, which become an integral part of the APM due to their strong implications in terms of parasitic torques and life capabilities.

Main design drivers for the APMs include

- Pointing accuracy, available to <20 arc-seconds
- Operational life, normally 15 years for telecommunication applications
- Stiffness, strongly linked to the payload inertia, but generally targeting frequencies above 2.5 Hz.

Crucial in the APM design is the compromise between stiffness and resistive torque due to the pre-load on the main bearings. Indeed a high bearing pre-load is desirable from the stiffness point of view and also to prevent gapping under vibration loads, but the high friction resistance torque limits the useful life. Developments in the direction of a variable bearing pre-load system are thus pursued in order to independently optimize the main bearing pre-load for the two different cases (stowed and operational).

Pointing ranges vary from a few degrees for fine pointing of telecommunication antennas, to about 12 degrees half-cone for inter-orbit link applications, to full 360° rotation for scientific missions.

Simulations of the entire dynamic chain, including APM and operated payload, are performed in order to avoid dynamic coupling and resonances between the structural elements and the APM generated forces/torques, which would be detrimental to the final pointing performance.

When stepper motors are used as actuators (as in the majority of cases) to limit APM generated micro-vibrations, techniques known as micro-stepping are implemented by means of a dedicated electronic design.

Operational temperatures ranging from approximately -40 °C to +60 °C are also critical for the design and sizing of inner components and particularly for the main bearing. Heat transfer through the APM is a concern because it generates temperature gradients through the main bearing, thus altering substantially the as-designed pre-load. A non-appropriate thermal design might cause: loss of stiffness, reduced lifetime, and severe malfunction due to excessive friction torque. A dedicated heater can limit the minimum temperature case (sometimes the motor self-heating capability is exploited for this purposes).

9.7.4 Hold-Down and Separation Systems

Most spacecraft have appendages (solar arrays, antenna reflectors, radiators, instruments, doors, sensors, booms, etc.) that are held stowed in order to fit into the launcher's available volume and to survive the launch loads and then deployed in orbit to their operating position.

Other equipment like scanning/refocusing mechanisms, electric propulsion pointing mechanisms or coarse pointing mechanisms must be stowed during launch and then released in order to allow in-orbit operations without any specific deployment.

To achieve these functions, two different types of mechanisms are used one after the other: the hold-down and release mechanisms (HDRM) and the deployment mechanisms (DM).

The hold down and release mechanisms are standard components for spacecraft in order to achieve mission-related critical functions. Their main functions are to secure during launch and to release once in orbit (or during descent to, or after landing on, a planetary surface) movable payload items, deployable appendages and separable mission elements. They can also be used to achieve timely synchronization for the deployment and/or ejection of specific appendages or separable mission elements.

The deployment mechanisms are used to enable deployment of a released appendage from its stowed position to its operational position by way of a defined kinematics and passive to active controlled dynamics. Once the final position is reached, the appendage is either latched at a defined position or the DM is used as a re-pointing or trimming device to achieve specific mission related functions. In some cases, HDRMs are not used in conjunction with DMs. This is, for instance, the case for scanning and refocusing mechanisms, electric propulsion pointing mechanisms and ejection mechanisms. Concerning HDRMs for spacecraft applications, they are generally composed of three functional elements

- A hold down preloading assembly (HDPA) such as a bolt, nut, threaded rod, tie-rod, cam/lever, cable, or rope which provides the required preload to be applied via manual operation or the use of specific ground support equipment in order to secure the equipment in stowed configuration during launch.
- Hold down release actuator (HDRA) which achieves the release of the preload upon the command of a drive electronic. The release actuator is generally mounted on the fixed part of the separable interface. It also frees and secures the separable interface from any mechanical links and, prevents any interference on deployment or operation of the appendage. For some applications, the bolt or threaded rod is ejected upon release and secured into a so-called bolt catcher via a dedicated spring and/or the stored strain energy.
- Hold down load carrying structure (HDLCS) which guarantees the launch loads transmissibility between the fixed part and the part to be released. This element completes the HDRM assembly and is always adapted to each appendage and spacecraft interface.

Table 9.4 HDRA technologies

Non-reusable	Partially reusable (need for refurbishment)	Reusable (manually resettable)	Reusable (self-resetting)
• Pyro cutters	• Pyro nuts	• Solenoid actuated nuts	• Electro-magnetic actuators and triggers
• Initiators	• Fusible wire actuated nuts	• SMA actuated nuts	• Magnetic clamps
• Pyrotechnic bolt, wire cutters and pyro-cutters.	• SMA direct actuators	• Paraffin actuators	
	• Spool based devices	• Wire triggers	
	• Separation nut	• Thermal cutters	
	• Thermal cutters		

HDRAs usually rely on one of the following technologies

- Pyrotechnic devices (release nuts/bolt cutter, separation nut, cutters, brazing melt, wire cutter, cable cutter)
- Split spool devices (fusible wire, SMA wires)
- Solenoid actuated nuts
- SMA triggered release nuts, with a temperature range of $-60\text{ }^{\circ}\text{C}/+70\text{ }^{\circ}\text{C}$
- SMA actuators (pin pullers/pushers) with a range of $-60\text{ }^{\circ}\text{C}/+70\text{ }^{\circ}\text{C}$
- Paraffin actuators (pin pullers/pushers) with a range of $-60\text{ }^{\circ}\text{C}/+80\text{ }^{\circ}\text{C}$
- Electro-magnetic/solenoid or piezo actuated pin puller/pusher actuators
- Electromagnets/magnetic clamps
- Thermal cutters/knife.

The HDRA can be located either in the load path or in a remote position and act via a cam/lever assembly. In this case, it is used as a trigger to initiate the release. Due to the generally external location of the HDRA, they must withstand a large temperature range, typically $-100\text{ }^{\circ}\text{C}/+120\text{ }^{\circ}\text{C}$ for most of the applications. Shape memory alloy and paraffin based technologies do not meet this temperature range, but can generate very low release shocks and can be used as triggers. It has to be mentioned that reusability is sometimes associated with lower reliability, as the number of parts in a reusable device is significantly higher than for a partially or non-reusable one. The underlying technologies can be grouped according to the level of reusability, which is a key feature with respect to HDRA implementation in space systems, as shown in Table 9.4.

In order to classify the different technologies, the HDRA could be divided into five categories with respect to their shock response spectrum peak upon operation

- High ($>3,000\text{ g}$)
- Medium (between 1,000 and 3,000 g)
- Low (between 300 and 1,000 g)
- Ultra-low ($<300\text{ g}$)
- No-shocks (indeterminately low).

In order to provide consistent, comparable, and recognized test data, the evaluation of release shocks requires approved and standard test procedures and facilities. For all HDRMs, the tightening tension can be settled and checked with different manners

- Flight torque or force sensors
- Tooling (on the ground) torque or force sensors
- Torque controlled screwing
- Angle controlled screwing.

Usually, the preload versus torque relationship is not well mastered by most of HDRM users. In most HDRMs, the actual tightening tension (minimal guaranteed preload), once the preload is applied, can hardly be known without the use of external force sensing devices, e.g. load cells or strain gauges. Users often rely on the preload versus torque relationship, which cannot guarantee a preload value accurate enough for space applications associated with the required repeatability. Historically, the high shock level generated by the release of the HDRA has been tolerated due to the following measures

- Definition of shock areas on the spacecraft where sensitive equipment must not be located.
- Equipment qualification to high shock levels.
- Damper implementation within the HDRM or at HDRM interface with the appendage and/or spacecraft.

However, as spacecraft have become more complex with an increased level of architecture modularity and versatility, a general trend has led to the reduction in the shock level generated by the HDRA itself. In addition, the trend away from pyrotechnic systems is growing because certain spacecraft do not allow pyrotechnics and substantial cost savings can be achieved by the avoidance of safety related costs. Another general trend is that telecommunication satellites have become bigger and heavier, while some science and Earth observation satellites feature composite architectures that require optical terminal mechanisms to operate under cryogenic environments. This has led to the need for a family of low-shock HDRA across the full tightening tension range [10 N–150 kN],

temperature range $[-130\text{ }^{\circ}\text{C}/+150\text{ }^{\circ}\text{C}]$, and at a competitive price. It has to be noted that the full tightening and temperature ranges will not be achieved with a single piece of hardware.

The use of shape memory alloys for HDRAs has been identified as a technology trend. However, its broad application for commercial applications can only be successful if their operating range can be increased up to $110\text{--}120\text{ }^{\circ}\text{C}$.

9.7.5 Position Sensors for Space Vehicle Mechanisms

In order to check or to control the position of space mechanisms, position sensors are necessary. Space mechanisms generally provide rotary movements and therefore require rotary position sensors. However, linear sensors are sometimes used. The following applies for both cases.

Position sensors can be characterized in terms of

- Performance, ranging from one position per turn to very high accuracy and resolution sensors (e.g. >24 bits per turn)
- Technologies
- Linear and rotary types
- Devices with mechanical contact or contactless
- Absolute and relative position signal.

All these position sensors are based on one of the following technologies

- Mechanical or electromechanical switches
- Electrical variable resistance sensors
- Magnetic/Hall effect sensors
- Inductive sensors (magnetic resolver, RVDT, LVDT, eddy current, Inductosyn, etc.)
- Capacitive sensors
- Optical sensors.

Position sensors are split into three categories linked to their performance level, which, in practice, result in the three following domains of applications

- A reference position sensor, providing one position per movement or one position per turn. These are usually named ‘switches’, and are in most cases for providing a TM (tele-measure) about the release and/or the achievement of a displacement/deployment. Switches are sometimes part of a closed loop to trigger safety mechanism power switch-off (heaters for actuators based on thermal phenomena: wax actuator, etc.).
- Sensors providing limited accuracy per linear movement or per turn. These are usually termed low and medium accuracy position sensors, or ‘potentiometer’ or ‘potentiometer equivalent’. Such designations result from the fact that people typically address resistive angular position sensors as potentiometers because most of these

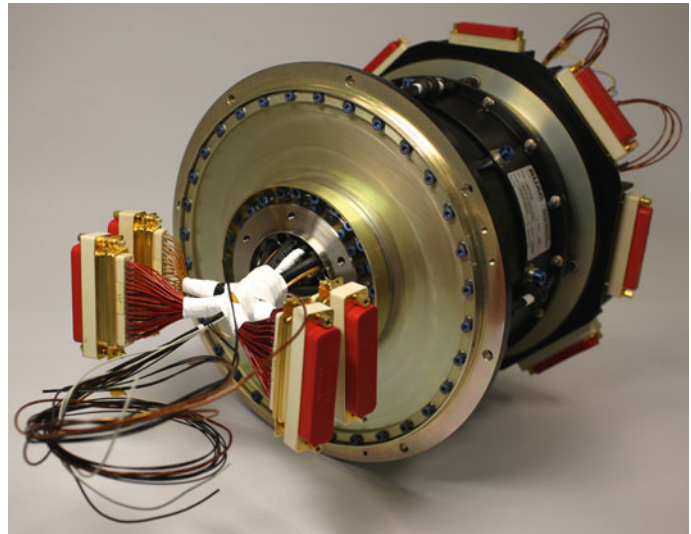
sensors are based on variable resistance techniques provided by means of a linear or rotating brush contact on a resistive path. These sensors are sensitive to the quality of the tribological contact between the wiper and the track, which can change with time and operation in a vacuum environment. This might change with the emergence of new and more reliable techniques. These new technologies, either magnetic, inductive, capacitive or optical ones, result in sensors with medium to high accuracy, with a cost dependent on performance. Apart from the low and medium accuracy criterion, this family of sensors is defined by its low cost and its low induced user constraints.

- High accuracy position sensors, often termed optical encoders because many of the applied techniques provide this level of performances using optical principles. For this category of sensors, the main requirements come from the extreme accuracy requested for scientific payloads (i.e. when a closed loop control of the position is required) but also for telecom equipment such as antenna deployment and pointing mechanisms. Apart from optical techniques, several other technologies can deliver these high performances. Further distinctions can be drawn between incremental encoders requiring a reference action upon each start-up, and absolute encoders that display the position at each power-on. Furthermore, there are also single-turn absolute encoders and multi-turn absolute encoders available. The latter usually incorporate a mechanical gear to register position across several 360° turns. Another distinction is ‘hard-coded’ bits that represent physical instances of, for example, the encoder glass disk, and electronically interpolated bits that are computed from the hard-coded bits and usually require extensive signal conditioning efforts within the control electronics loop.

Switches and potentiometers are low cost position sensors. Optical encoders are more expensive, with the cost being performance-dependent. For high accuracy sensors, special attention should be given to the variety of units that are used to designate an angle (bits, degree, part of degree, arc-minute, arc-second, micro-rad, etc.). Although the rotary position sensor aims to provide an angular position (from 0 to 360°), for which the more common unit is the angular degree, the performance of a high-resolution angular position sensor is often expressed in bits or in arc-seconds. Each of the high position sensor performances can be quantified by any of these units, although it is common to use ‘bits’ when speaking about resolution and arc-seconds when speaking about accuracy and repeatability.

An optical encoder is used most of the time in the closed control loop of a mechanism, especially for very accurate pointing or scanning. In most cases, advanced filter and signal treatment are built into the control loop in order to obtain extreme performance. This signal treatment is often

Fig. 9.31 Solar array drive mechanism manufactured by RUAG Aerospace Zurich



numerical, which requires a level of resolution significantly higher than the level of accuracy provided by the position sensor. Therefore, it is necessary that the optical encoder resolution be much higher than its stated accuracy, in order to get an insignificant impact of the computation error on the overall performance.

9.7.6 Solar Array Drive Mechanism

Most three-axis stabilized satellites use solar panels to generate electrical power for their equipment. For the best performance, these panels must be aligned perpendicular to the Sun. In most cases, when the satellite's body is pointing towards a defined target and its orientation is not fixed with respect to the Sun, a relative motion between the satellite's body and the solar panels must be provided. The rotating mechanism performing this task is the solar array drive mechanism (SADM). A motor is used to rotate the solar array at the required speed and in the required direction, and a specific electrical device (slip-ring, cable-wrap, twisted capsule, et cetera) is used to transfer the power (and data) between the solar array and the platform. The SADM is one of the most critical hardware components of a spacecraft (i.e. it is a single-point failure) and its design is usually optimized with respect to the specific satellite platform and its power needs, which can often cover a range from ~ 500 W to $\sim 20,000$ W per solar array wing.

Typical main and secondary functions of a SADM include

- To mechanically link the solar array to the satellite, whilst allowing the solar array to rotate around a specific axis.
- To rotate the solar panel via the solar array deployment electronics (SADE) or in response to a satellite command

to maintain the solar array pointing at the Sun, to rotate the solar array into a reference position, to rotate the solar array in a high-speed mode, to maintain the solar array in a defined/fixed position, and to provide telemetry signals from the SADM (including the SADM angular position sensor).

- To transfer the solar array electrical power to the satellite.
- To transfer signals and low power lines between the 'rotating' part (solar array side) and the 'fixed' part of the satellite (platform).
- To assure the solar array grounding to the satellite.

Driven by these functions, a SADM consists of three major subsystems

- The rotary actuator
- The electrical transfer unit
- The angular position and reference sensors (Fig. 9.31).

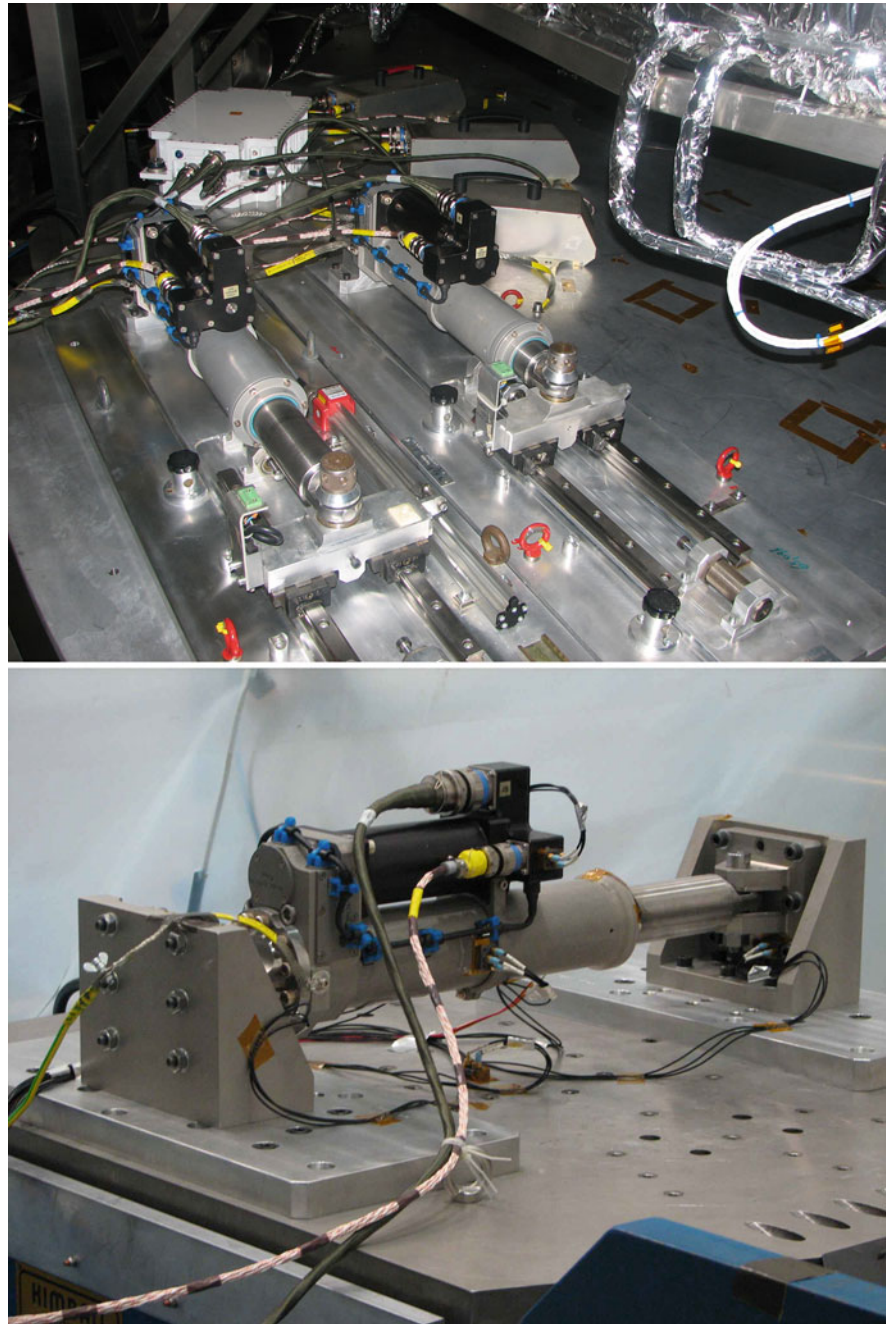
9.8 Mechanisms for Launchers

9.8.1 Electro Mechanical Actuators for Thrust Vector Control of Launchers

The thrust vector controller (TVC) is a subsystem that controls the direction of gimbaled nozzles of rocket engines in response to a request/command from the launcher trajectory and attitude control system. The subsystem comprises three major items

- Battery or batteries (see Chap. 10 for an in-depth discussion)
- Control unit with a dedicated controller and drive electronics to command and control the mechanical actuator linear displacement

Fig. 9.32 Vega ZEFIRO electro mechanical actuator undergoing an extended performance test aimed at the verification of the subsystem operational envelope (*top*) and undergoing vibration testing (*bottom*). Image S.A.B.C.A



- Linear actuators (in pairs, acting on 90° planes) which can be operated either with a hydraulic system (as for the Ariane 5 main engine) or with an electrical motor (as is the case for the four stages of the Vega launcher); in the latter case the actuator is called electro mechanical actuator (EMA).

All three items are connected by a cable harness for the transmission of power and data signals. EMAs have become attractive candidates to replace hydraulic actuators for thrust vector control, thanks to an easier implementation and lower maintenance requirements. The EMA is a

mechanism composed of an electrical motor (either co-axial or parallel to the piston) connected to a gearbox that drives a roller screw. The rotational motion of the electrical motor is transformed into the linear motion of a piston via the roller screw. The linear position of the piston is measured via a resolver connected to the electrical motor and a linear variable differential transformer (LVDT) sensor, one part of which slides inside the roller screw and is rigidly connected to the piston and follows the motion. Figure 9.32 shows the Vega Zefiro TVC subsystem being subjected to an extended performance test and to vibration testing. Due to high

dynamic transient loads acting on the EMA, the roller screw is subjected to very high rotational speeds and this represents a physical limitation together with the electrical power to drive the motor since it can lead to unacceptable EMA dimensions and mass. The EMA's function is not only to act as a directional device for the launcher but also to absorb shock loads generated by liquid fuel rocket engines at their ignition as a result of initial combustion instability. In either case, the performance of the EMA is strongly coupled to the launcher dynamics and its control system, and it ultimately influences the loads transmitted to the payload by the launcher.

9.8.2 Launcher Valves

There are numerous flow control valves in the main and upper stages of launchers with liquid-propellant rocket engines. The application of valves range from cryogenic, e.g. liquid hydrogen, chamber flow control, operating at 20 K, up to hot gas applications like turbopump feed control valve at 1,000 K. They operate under pressures as high as 260 bars. Apart from the extreme operational temperature ranges, there are requirements for leak tightness and precise regulation of the fluid. In addition to simple on/off switching applications there are proportional flow control valves of different sizes and speeds.

The configurations include butterfly valves, poppet valves, ball valves, grid valves for fluid control as well as mono-stable or bi-stable on/off valves, and they can be actuated pneumatically, hydraulically, or electrically.

There is a trend to simplify and optimize power distribution and energy usage within launcher engines. This involves the replacement of hydraulic, pneumatic, and electric components that require multiple generation sources, energy conversion methods and redundant lines with a single electrical architecture. Fluid components such as pipes, tanks, accumulators and valves are replaced with electrical harnesses that require little or no maintenance, are simple to route, and are significantly more robust and damage tolerant than their fluid equivalents. Power generation and the control of energy distribution are also simplified. Electrically actuated valves are equipped with (redundant) actuators, gearboxes, and position sensors.

9.9 Tribology

Every space mechanism has moving parts. Most of the time, these are in contact with fixed parts. To reduce friction and wear effects as much as possible, lubricants are used between the parts in contact. The science that studies the phenomena of friction and wear, including lubricants,

their characteristics and interaction with the parts of the mechanism, is known as tribology. Three main types of lubricants are used for space mechanisms. Note that oils and greases are often grouped as liquid lubricants.

9.9.1 Oils

Oils are very effective lubricants at sufficiently high speeds because they tend to create rather stable friction conditions and resistive forces/torques (i.e. low noise). However, they tend to evaporate and to migrate by surface creep. Therefore oils are especially suitable for mechanisms that are sealed, for example reaction wheels. It is important when designing a mechanism to take into consideration degradation of the oil, particularly if using perfluorinated polyether (PFPE) oils like Fomblin Z25. The fluor contained in these oils tends to react chemically with the iron of the steel in the contacting, moving parts (for example, bearing balls and races). This leads to a breakdown of the oil, which is detected by an increase in the contact loads or torques. This increase of contact loads is due to a polymerization of the oil (creation of solid particles). Some oils like mapping and comparing (MAC) oils (for example Nye2001a) do not contain PFPE and are therefore less sensitive to lubricant breakdown.

9.9.2 Greases

Greases are widely used in space mechanisms because they tend to generate relatively low noise and provide good lubrication even at low speed. They are also less sensitive to evaporation and surface creep than oils. In fact grease for space applications is oil (known as the base oil of the grease) in which solid substances are added to assist the lubrication process, for example, polytetrafluoroethylene (PTFE) particles. The same problem with the degradation of the lubricant exists for the greases, as grease contains its base oil.

9.9.3 Solid Lubricants

Solid lubricants tend to generate more noise in the resistive forces and torques but do not evaporate. They are therefore suitable for extreme temperatures (cryogenics, or very hot applications) or in applications where contamination by condensation could be an issue (for example, in optical systems). Solid lubricants will be used, for example, on BepiColombo where temperatures of about 250 °C are expected. The most used solid lubricants for space mechanisms are sputtered molybdenum disulfide (MoS₂) and ion plated lead. Sputtered MoS₂ exhibits a lower friction than

the ion plated lead (and it is typically less noisy) but has a lower life expectancy. Neither lubricant is suitable for use in air, as the friction coefficient significantly increases and the life expectancy is considerably reduced.

To use solid lubricants during ground testing in air (which has the advantage of leading to reduced cost), the solid lubricant can be located in a reservoir which, in case of a bearing, is the cage. In this case, the lubricant is transferred from the cage to the balls and from the balls to the races. Thanks to the reservoir, even if the solid lubricant is degraded in air, new fresh lubricant is provided by the reservoir. For sputtered MoS₂, the reservoir (or the cage in case of a bearing) can be made of special compound materials, e.g. PGM-HT (a composite of PTFE, glass fiber, and MoS₂). For ion plated lead, the reservoir (cage in case of a bearing) can be made of leaded bronze. To help with lubrication before the lubricant transfer is installed, sputtered MoS₂ will be also placed on the contacting surfaces (balls and races in case of a ball bearing using a PGM-HT cage). When using a leaded bronze cage, ion plated lead will also be placed on races for the same reason. It is important to perform the run-in of the mechanism (or the bearing) in vacuum because of the degradation of the sputtered MoS₂ and ion plated lead in air. A typical run-in for a bearing consists of about 50,000 revolutions. It is to be noted that while PGM-HT can withstand several million revolutions in air after run-in, leaded bronze is limited to not more than 100,000 revolutions. Both PGM-HT and leaded bronze can provide life expectancies of several hundreds of million revolutions in vacuum in the case of a bearing.

9.10 Design and Verification of Space Vehicle Mechanisms

9.10.1 Numerical Simulations: Multi-body Dynamic Simulation

In recent years, space-system design has shown a clear trend towards increasingly complex configurations. Typical examples are the use of several flexible components (antennas and solar arrays), the need for deployment and retrieval mechanisms, the demand for high precision pointing systems, and an increase in mission scenarios that involve the assembly of large structures in space. This trend has also caused an evolution towards a multi-disciplinary design approach, particularly in the area of dynamics and control.

In order to study the performance of generic controlled dynamic systems, it is essential to have a dedicated tool that allows the user to model, in a short time, the complex behavior of the system's dynamics, and their interactions with the control. In fact, some systems require a model with

more than one body in order to take into account their different characteristics and their mutual dynamic interactions. This is a non-simple task, requiring time to understand, code, and validate the dynamic behavior of the system. A large amount of research has gone into the development and improvement of multi-body software, with the aim of reducing the time to model a system and the computation time required to run an analysis. Multi-body software involves the derivation of the equations of motion for multi-body systems, which are systems characterized by several bodies connected by hinges that permit relative motion across them. Based on the latest improvements in software and technical experience, the modeling and simulation approach enhances the design and verification process of aerospace mechanisms/systems. This refers to several aspects such as

- Anticipate and understand hardware performance
- Identify criticalities of key design parameters by means of parametric/sensitivity analyses
- Optimize the hardware design
- Interpret the hardware performance by a model-assisted approach
- Diagnose any potential anomalies.

Typical multi-body dynamics activities are performed in support of numerous projects

- Mechanisms
 - Design and analyze mechanisms systems
 - Assess parametric design solutions and to optimize the performance of structures and mechanisms
 - Predict the static, kinematic, and dynamic behavior of mechanical systems
 - Animate the motion of dynamic, solid models
 - Simulate control loops incorporating mechanical parts, e.g. active structures of mechanisms
 - Perform conceptual design studies and correlate test and analysis data
 - Investigate the performance and possible malfunction of mechanical systems, including in orbit
 - Modeling complex devices and mechanisms control design
 - Coupled system frequencies and time responses.
- Spacecraft
 - Modeling unconventional spacecraft dynamics, including orbital environment disturbances,
 - Control laws, including sensors and actuation dynamics,
 - Coupled system frequencies and time responses,
 - Docking phase analyses.
- Launch Vehicles
 - Non-linear time dynamic simulations for flexible bodies with time varying characteristics
 - Atmospheric environment including external disturbance
 - Launch vehicle flight dynamics-control interaction

- Frequency domain analyses
- Multiple nozzles dynamics capabilities
- Local analyses (gust response, lift off, multi-payload separations)
- Collision avoidance
- Data recovery procedures.

9.10.2 Electromagnetic Simulations

The rapid development of simulation tools and computing capacity has made finite element (FE) analysis a very useful tool for mechanical design verification. This trend is also confirmed in the field of electromagnetic simulation (and multi-physics simulation in general), where several software houses are developing tools for the design and analysis of 3D/2D problems, such as motors, actuators, transformers, and other electrical and electromechanical devices that are common to automotive, military/aerospace and industrial systems. In particular, for space applications a big benefit can be gained by using these tools to design and optimize devices such as

- Electric motors (stepper, brushless DC, etc.)
- Custom-design rotary and linear actuators
- Active and passive magnetic bearings
- Electromagnetic and capacitive sensors
- Contactless power and signal transfer devices
- Electromagnetic brakes
- Hold-down and release devices (based on electromagnetic operating principles)
- Special systems.

The software packages most commonly used to perform these analyses are various. The French company CEDRAT has developed the commercial software package FLUX[®], which is widely used by European industries. In global terms, Maxwell[®], ANSYS[®] Multiphysics Solutions and COMSOL Multiphysics[®] are widely used, primarily thanks to their ability to interface with a multi-physics environment for coupled thermal, mechanical, and electromagnetic analyses. Furthermore, open source software is available for certain problems, for instance Finite Element Method Magnetics (FEMM) for 2D magnetostatic simulation. In general, the capabilities of these tools allow the study of

- 2D and 3D problems
- Static and transient analyses
- Coupling with mechanical (rotating and translating) motion
- Coupling with external electrical circuits
- Solution of electric field problems
- Coupling with external mechanical loads
- Interface with MATLAB[®]/Simulink[®].

9.10.3 Verification Tests of Mechanisms

The required reliability of a space mechanism is much higher than is usually necessary for general ground mechanisms, primarily because it is usually impractical to implement repairs during its operation. The capability of the mechanism to fulfill its requirements as well as to maintain the desired performance throughout the entire mission, without repair, must therefore be assured. In order to achieve this, a rigorous requirement verification process must be adopted. The verification process involves demonstrating requirement compliance by means of one or more of the following methods: review of design, analysis, inspection and test.

Verification by test is implemented on a set of selected models chosen for the project. This ‘model philosophy’ is defined by means of an iterative process that combines programmatic constraints, verification strategies, and the integration and test program, taking into account the development status of the candidate design solution. The first method consists of performing a verification of the design documents, reports, and technical description in order to prove that a requirement is met unambiguously. The second method employs accepted analytical techniques to provide evidence of a requirement’s fulfillment. Verification by inspection is achieved by visual determination of a characteristic (e.g. construction features, presence of an element). The test verification of a requirement consist of experimentally measuring one or more parameters or functions within a representative environment.

9.11 Space Vehicle Mechanisms Materials and Processes

Rules for selecting materials to build spacecraft mechanisms follow general materials selection rules. Regarding their functionality, the selection of materials is a trade-off process that takes into account the following

- Mechanical properties (e.g. strength, stiffness, fracture toughness, fatigue resistance, micro-yielding, creep)
- Physical properties (e.g. coefficient of thermal expansion, coefficient of moisture expansion, thermal conductivity, electrical conductivity, thermo-optical properties)
- Chemical properties (e.g. corrosion, susceptibility to hydrogen embrittlement)
- Interfacial characteristics (e.g. mechanical contact surface effects such as self-lubricating capabilities, susceptibility to cold welding or galling)
- Combinations of the above (e.g. stress corrosion, corrosion fatigue).

Regarding the mission constraints, the material selection must take into consideration the specific environments on the ground, during launch, and in space. This includes

- Temperature and thermo-cycling
- Vacuum (outgassing)
- Radiation
- Electrical charge and discharge
- Fluid compatibility
- Galvanic compatibility
- Atomic oxygen
- Moisture (absorption and desorption)
- Crewed environment (for hazard and risk potential, both structural and physiological) additional requirements such as outgassing, toxicity and odor, bacteria and fungus growth, flammability, etc.

Regarding the space mechanisms materials interfaces, many requirements are related to the surface characteristics and the most adequate surface treatment must be applied in 'ad equation' with each specific application's needs. The surface treatment aims in general to

- Protect the surfaces against corrosion, cold-welding, fretting, lubricant creep
- Improve electrical or thermal conductivity and thermal-optical properties
- Lubricate the surface and improving its wear resistance
- Improve the tribological-surface quality by polishing, cleaning, or hardening.

Based on these factors, the main materials used in mechanisms are aluminum alloys (2024 T8, 6061 T6, 7075 T73, etc.), stainless steel (300 series, 400 series including 440C for ball bearings, 15-5 PH with H1000 and above, 17-7 PH CH900, etc.), nickel alloys (Inconel 718, etc.) and titanium alloys (Ti6AlV, etc.).

9.12 Deployable Structures for Space Applications

Envelope limitation within launcher fairings has always been a major design driver for spacecraft configuration; sometimes, if not often, more severe than the mass limitations. Various techniques to reduce the spacecraft volume at launch have been developed, giving rise to the entirely (and almost space specific) topic of 'deployable structures Space technology'. A significant number and type of deployable structures have been successfully operated in space, ranging from very small booms to large antenna reflectors. Basic building blocks of a deployable structure include

- Joints, to provide relative motion
- Actuators, to provide deployment energy

- Hold-downs release mechanisms, to keep the structure in its stowed state during launch
- Dampers, to reduce end of travel shock, when energy storage based mechanisms are used
- Latches, to ensure that the structure locks in its final deployed configuration
- Interconnecting rigid structural members (struts, panels), to provide structural support to the payload to be deployed.

Although deployable structures such as magnetometer deployable booms, thermal shields, and articulated masts are common, it is often possible to observe a less distinct separation between the different elements, particularly with structural members that also provide deployment energy (e.g. by deformation in tape spring mechanisms) or trajectory guidance. Indeed, an entire class of 'self-deploying' structures can be formed by applying pre-load or deformation to specific structural members.

Generally, the main concern for the deployable structure designer is that the deployment function is successfully achieved, with the required end-of-deployment performance. Due to the extreme difficulty of perfectly reproducing on the ground the same environmental conditions to be experienced in orbit (and specifically the impossibility of perfectly reproducing the absence of gravity, especially for large structures), the final verification of deployment functionality will always be associated with a feeling of imponderability and generate major relief when successfully achieved in operational conditions. To minimize the intrinsic risk of any kind of deployment malfunction, the designer is forced to employ concepts that are as reliable as possible, which in turns means as simple as possible, to achieve an optimal synthesis between the various requirements and constraints. This effort is normally rewarded by the elaboration of design concepts that can be said to be 'elegant'. While this term is difficult to technically define, it is generally well recognized by peers and customers, and sometime can be an important aspect of a commercial success. It has also to be noted that deployable structures have to be considered, and approached, as a system level issue. Indeed, even in the simple case of short deployable booms, their number, location/orientation, both in the stowed and deployed configurations, have an impact at system level (occultation, risk of collisions, mechanical interference, and thermal fluxes implications). And for large deployable structures, the overall spacecraft configuration and attitude control subsystem might be driven by the deployable structure's architecture and performance.

9.13 Design drivers for Deployable Structures

9.13.1 Deployment Reliability

The main design driver for a deployable structure is, in almost all applications, its release and deployment reliability. Despite such fundamental relevance, few requirements are more difficult to quantify than those relating to reliability. Techniques and approaches derived from electronic components are employed, based on standard reliability figures for elementary parts/couplings. A statistical database for the reliability figures applicable to the real structure and components under design are typically of limited availability, and above all, the overall reliability of a deployable structure can rarely be defined by the summation of individual reliability parts.

Reliability concepts have to be incorporated into the design in its very first stages, by means of a balanced and harmonized design approach. To exploit hardware or company heritage in components and concepts is natural, but innovation must not be disregarded. Redundancy must be implemented, but not be intended as simple duplication of parts or a means of dealing with complexity. Simplicity of design solutions is a must, but should not be confused with simplistic approaches. Robustness must be pursued, but without overdesigning. Overall, a deployable structure operates as a unique organism, where individual components harmonically cooperate for a successful performance of the entire system.

9.13.2 Stiffness

Stiffness requirements are normally applicable and critical for deployable structures. Typical stiffness requirements for deployable antenna reflectors are in the range of 2.5 Hz and above. Larger structures, like solar arrays or large reflectors, can have first eigenfrequencies below 1.0 Hz (even down to few tenths of a Hertz in some cases). Although it might seem that the stiffness requirement affects mainly the static/structural parts of the mechanisms (e.g. the hinge shaft section or the mounting flange sizing), it has to be stressed that it might strongly influence the essential ‘mechanism’ aspect of the design. Indeed a higher or lower stiffness requirement might imply the need (or not) for a latching system, which may alter substantially the overall mechanisms design or its components selection. Similarly, a higher or lower first eigenfrequency might imply higher or lower end-of-travel shocks, and the need (or not) for a

damping device (see the following [Sect. 9.13.4](#) on actuation margins). It has also to be noted that the stiffness requirement associated with the deployable structure is often attributed on the basis of ‘rigid spacecraft mounting interfaces’. The degree of validity of this assumption is often questionable, and sometimes it affects the first frequency of the deployed structure by a factor of ten. It is therefore essential to establish a dialogue with system engineers to avoid over-designing critical parts of the mechanisms, which besides other important negative effects might result in a less reliable deployable structure.

9.13.3 Accuracy/Stability

Accuracy of the final deployed configuration is sometime an essential requirement (e.g. high-frequency antenna reflectors, optical telescopes, and so forth). Accuracy can be in the range of several thousandths of a degree (or less). High accuracy normally requires the implementation of a latching device, possibly of a type that generates some degree of pre-load into the system, once actuated. By pre-loading the joints, residual backlash is recovered and the final deployed structure will have better repeatability and superior accuracy performance.

9.13.4 Actuation Margins

As any mechanism-operated system, deployable structures have to show adequate margins versus actuation forces/torques against opposing resistance forces/torques. One complication is the presence of a harness routed across its joints and it is important to define/characterize the parasitic effect at an early stage of the design. In addition, thermal effects (low temperatures in particular) can fundamentally influence the level of parasitic effects. An easy solution would be to incorporate enough margins in the design to account for poor assumptions. Unfortunately, excessive actuation margins may result in severe risk of structural damage at end of deployment (or during deployment). For example, for energy storage based actuation such as spring actuated deployment, the end-of-travel shock may be a design driving case. This is particularly true when thermal effects reduce friction-resistance components. A speed regulator or damping device might then be deemed necessary, substantially affecting the overall reliability of the mechanism. Again, a balanced design (in this case avoiding excessive actuation torque) is essential in order to maximize overall reliability.

9.14 Verification of Deployable Structures

The on-ground test verification of deployable structures is strongly affected by the presence of the gravity field. Gravity off-loading jigs are normally necessary, with few exclusions. Testing under one 'g' conditions is further complicated by the fact that off-loading jigs should be compatible with thermal vacuum chambers in order to perform functional testing under representative environmental conditions. Obviously, making an off-loading system compatible with extreme temperatures and vacuum is a challenging and very expensive task. Furthermore, thermal vacuum chamber volumes are typically quite limited, and off-loading devices generally large (both in height and in-plane). Hence, verification in thermal vacuum chambers of full deployment functionality, particularly for large structures, is seldom performed. Alternative qualification logics are based on qualification at different levels (equipment, subassembly, etc.) and limiting the thermal vacuum test to a partial deployment (release from hold-down points with limited angular motion). For a large structure, like an antenna reflector, in-orbit validation would be required in order to commercialize the product.

In order to support the design phases, and to complement the on-ground verification phases of deployable structures, simulation activities have come into play. Indeed, for large solar arrays, with multiple hinge axes oriented in different orthogonal directions, even room temperature off-loading systems are almost impossible to conceive.

9.14.1 Simulations

Space-system design has shown a clear trend towards increased configuration complexity in response to challenging mission requirements. The need for large apertures has been one major driver in this respect, including large solar arrays for increased power generation, large antenna reflectors, large sunshields for scientific applications (e.g. JWST) and solar sails. Common to those applications are the use of several flexible components, the need for deployment and retrieval mechanisms, and above all the need to verify on the ground the deployment functionality and the overall system performance. Representative deployment and performance testing on ground of such large structures are often very difficult if not impossible (mainly due to gravity, but also including air and other disturbing effects). Consequently, the need for increasingly sophisticated simulation tools and techniques has emerged. This trend has also caused an evolution towards a multi-disciplinary design/verification approach, with its major emphasis in the area of dynamics and control (particularly

as increased accuracy and stability performances are increasingly sought for pointing payloads and antennas).

Among the broad number of deployment applications, there are three main classes of problem, as discussed below.

9.14.1.1 'Conventional' Deployment Systems

These systems are characterized by mainly rigid motion among the different mechanical elements, taking into account small deformations due to the structural dynamics. Furthermore, non-linear hinge characteristics (friction, backlash, hysteresis, etc.) are also taken into account during the analysis. The main requirement is to properly model lightweight structures with variable boundary conditions due to the presence of motors at the hinges, locking systems, friction, and so forth. Advanced multi-body software techniques permit the investigation of coupled rigid and structural dynamics with a control system, aiming at pointing accuracy/stability performance verification.

9.14.1.2 Deployment Systems with Large Deformation

For applications like large deployable antennas or some solar sails, it is not realistic to neglect effects due to large deformations of the mechanical parts. For this class of problem, either multi-body software capable of simulating the non-linear behavior of some specific structural elements, or finite element codes that take into account the rigid body dynamics, are currently used.

9.14.1.3 Deployment of Inflatable Structures

The last class includes the deployment of inflatable structures, where it is important to simulate the inflation dynamics, the wrinkling of thin membrane, the definition of the initial shape, the fluid structure interaction, etc. The simulation of inflatable structures is a very demanding task. This is essentially due to the highly non-linear nature of the involved phenomena, including large variation in shape and in mass density. Moreover, complicated physical phenomena, such as wrinkling of a very thin membrane might occur. The solution of the governing equations of the problem is achieved at a very high computational cost.

9.15 Categories of Deployable Structures

9.15.1 Single Deployment Appendages

One-shot deployable appendages typically include booms carrying magnetometers or other small payloads, antenna reflectors (up to about 3 m diameter), solar radiators, thermal shields, and a variety of small panels.

The typical alternative approach is whether to use a motorized deployment or one based on energy storage such as a preloaded spring. Depending on a case-by-case basis, the optimal solution is selected, with the motorized option being preferred. The case where repeated deployment/stowage are required is excluded here, as only an electrical motor based solution is realistically possible. Indeed, when a substantial amount of harness is present, and in particular when deployment occurs at a low temperature (below $-20\text{ }^{\circ}\text{C}$), a large reserve of actuation torque has to be foreseen. This is normally not feasible with spring driven hinges, where a torque excess would result in high end-of-deployment shock. That is one main factor for the selection of an electrically actuated deployment mechanisms.

As previously noted, the end-of-travel shock is a driving requirement. To reduce its level, speed regulators or dampers are used. Those devices are based on a number of different working principles such as dry friction, viscous fluid effects, parasitic eddy currents, or more exotic concepts like low temperature melting alloys or clockwork escape mechanisms.

Concerning the deployed accuracy, adjustable end stops are typically implemented, with the contact surface shapes rounded to minimize Hertzian contact stress and to maximize accuracy and repeatability. Constant-torque types of spring are also implemented to realize a smooth deployment, with minimal variation of the actuation torque. Maximum care has to be taken in the design of the harness routing across the hinge unit. This is fundamental in order to minimize (and above all ensure repeatable) harness induced parasitic effects. Also the design of the hinge's thermal protection hardware, normally in the form of multi-layer insulation blanket (MLI), is very important, since a non-optimal MLI lay-out or fixation system may result either in hot-spots (where Sun-trapping occurs) or in the mechanism jamming at deployment. The minimum distances between fixed and mobile MLI layers is specified in the European Mechanisms Standard is 15 mm.

9.15.2 Uni-dimensional Deployable Structures

So-called uni-dimensional deployable structures represented by the booms and masts of many mission applications. Major technology types include

- Rolled, open-section tubes based on metallic thin foils, like the Storable Tubular Extendable Member (STEM) and 'Bi-STEM' types used to deploy the Hubble Space Telescope solar arrays
- Rolled, closed-section tubes, like the collapsible tube mast (CTM) developed by SENER in 1986, and more recently the collapsible CFRP boom developed by DLR

- Coilable masts, like the CoilABLE Mast from the ABLE company in the US
- Articulated foldable masts, like the ADAM and FAST mast, also from the ABLE company, being used for ISS solar panel deployment and for the NASA/NGA Shuttle Radar Topography Mission (SRTM) mission
- Telescopic masts, in different size and technologies, but always limited to a few meters length, and with less impressive stowed-to-deployed length and linear mass values.

The main differences between these technologies are mainly in the stiffness, which, not surprisingly, is inversely proportional to the mass per unit length. In particular the first two categories (open and closed section rolled tubes) are giving the lowest specific mass (mass per unit length) and package efficiency (ratio between stowed and deployed length), whilst the coilable and articulated masts provide the best stiffness performances.

Telescopic booms form a particular category. They cannot compete with the specific mass of other categories, nor with their superior maximum extension capabilities (up to 60 m for the ADAM mast, in-orbit), but still can compete for shorter lengths with a high stiffness requirement. It has to be noted that the longer and stiffer the application, the more complex the boom structure becomes, going from a simple thin foil of metal or CFRP for collapsible booms, to the very complex truss type of structure for the ADAM and FAST mast.

Concerning the final as-deployed accuracy of booms, milli-metric requirements are applicable for length of 12 m and more. Consequently, a backlash-free design is needed. This is easy for the monolithic rollable-collapsible tube concepts, but requires high design skill for a truss where sequential locking of individual bays must be ensured for overall truss mechanical performance. Low coefficient of thermal expansion materials (CFRP) are used whenever possible in order to ensure an overall high thermo-elastic dimensional stability of the deployed structure. Cross section size becomes particularly important for very long applications, where the overall buckling load (due to mechanical load, but enhanced by thermal effects) might be the driving requirement.

9.15.3 Bi-dimensional Deployable Structures

Bi-dimensional deployable structures typically include

- Large antenna reflectors
- Large thermal shields
- Solar arrays
- Solar sails
- Lenses for solar concentration.

Large solar arrays are a mature technology, but the state-of-the-art is still progressing due to increasing array surface requirements. Associated mechanisms include classical components (spring or electric motor based), advanced components (for example, tape-spring based inter-panel hinges), and inter-panel synchronization systems. For this class of deployable structures (as well as for large thermal shields) verification of the final performance (particularly in the relevant environment) is a critical issue, and advanced simulation techniques (as discussed earlier) are becoming increasingly important for the final product qualification.

Solar sails and large solar concentrators are still (with very rare exceptions) at the laboratory prototype level. Due to the very large dimensions (sometimes spanning 50 m and more), suitable step-by-step development approaches and intermediate-size validation models for flight verification purposes are proposed. Simulation capabilities are also essential for design consolidation and validation.

9.15.4 Special Cases of Deployable Structures

9.15.4.1 Large Deployable Antenna Reflectors Requirements

Antenna reflectors belong to the category of external spacecraft appendices that are most severely exposed to the thermal, radiation, and mechanical environments, both during launch and in orbit. In addition, the requirements for radio frequency performance, dimensional accuracy and stability, mass, and stiffness drive the design and verification. The only solution to these demands is to use the most advanced materials, manufacturing technologies, and engineering methods. The case of large antenna reflectors is a special one due to the dimensions (typically above 4 m projected aperture) since the antenna is ‘built’ in space following a rather complex sequence of release, deployment, tensioning, and locking steps. Each of these operations can represent a single-point failure for the mission. This section provides a general classification of large deployable reflectors based on the technical solutions available, a brief summary of existing products and their performances, critical technologies, and verification methods. Considering telecommunications, Earth observation, and scientific applications, the functional requirements can be divided into two main groups. In the first group, the functional requirements result in a projected aperture of 4–7 m with a root-mean-square (RMS) surface accuracy of around $\lambda/50$ (i.e. 2.5 mm in S band) for applications mostly in L and S bands or $\lambda/50$ to $\lambda/100$ for applications from C to Ka band (i.e. 0.1 mm in Ka band). The second group corresponds to projected apertures in the range of 9–25 m with an RMS surface accuracy of around $\lambda/50$ for applications mostly in UHF and L/S bands. The pointing stability of the

reflector with an arm attached should be in the range of 0.020–0.1 °half-cone. Concerning mass, the technology state-of-the-art is around 1 kg/m² for the very large apertures, but for higher accuracies 2 kg/m² is often necessary.

9.15.4.2 Deployable Reflector Types, Technical Solutions and Existing Products

Large deployable reflectors can be classified according to the reflecting surface technology and the supporting structure architecture as summarized in Table 9.5, along with their realization as flight products or advanced development models. This classification and product list is not exhaustive, it merely gives a sense of the technology.

The most numerous class is the metal-mesh reflecting surface, for both historical reasons and technical maturity, especially for the largest dimensions. Metal meshes require tensioning in order to acquire the parabolic shape and to generate a stable electrical contact between the metal wires. Their shape accuracy is a function of the facet dimensions or number of attachment points, as well as the tension applied. The materials of the mesh (typically molybdenum tungsten wires) and the backing structure (CFRP and aramidic fibers) control the thermal stability. Among the architectures available, the peripheral ring structure and its variants offer the best performances within the diameter range of about 6–15 m. Meanwhile, a modular construction can cover the full range of dimensions (Fig. 9.33).

Membrane technologies have been employed in inflatable structures, as well as in mechanically tensioned surfaces. The well-known thin metalized polyimide films or aramidic fiber fabrics require tension in order to stiffen and stabilize the shape. Despite remarkable progress, in-orbit polymerization and rigidization of inflatable structures has not yet been commercially adopted for the demanding RF requirements mentioned.

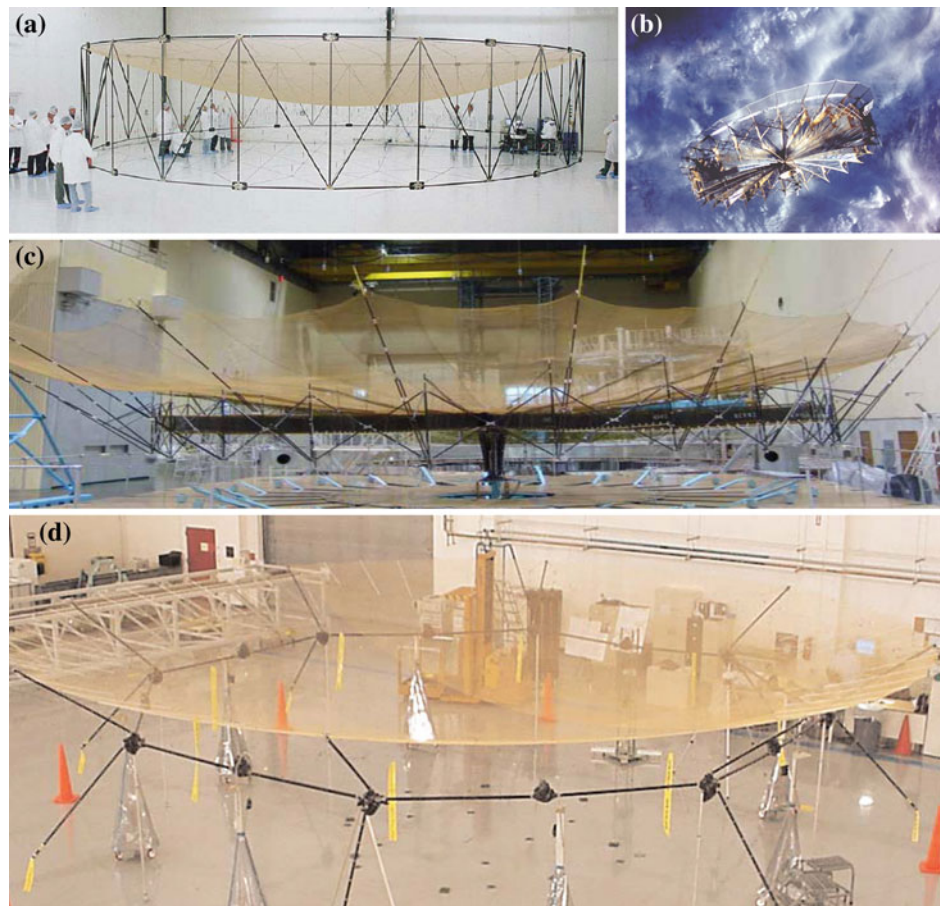
As regards the mechanical tensioning of membranes, surface accuracy has been achieved on flat surfaces for low RF frequency operation (typically P to L bands), e.g. for multi-layer SAR sub-arrays. In the case of doubly curved surfaces, the technology applied for mechanical tensioning of the membranes often results in demanding compromises of surface accuracy, deployment reliability, and mass.

Deformable shell reflectors (also known as ‘spring-back’) offer a lightweight alternative solution when folding in only one direction is sufficient. Deployment motorization is provided by the elastic energy stored in the deformed shell. Due to the use of CFRP, often made of triaxial fabrics and epoxy resins, the strength and creep of the fibers and resin limit the deformed curvature. The maximum dimension achievable by this class is typically a projected aperture of 6–7 m due to the folding. An alternative is to use silicone as the matrix, to achieve more efficient packaging. The

Table 9.5 Classification of large deployable reflectors

Reflecting surface technology	Supporting structure architecture	Realizations
Metal mesh	Peripheral unfolding ring structure	Astromesh AM-1, AM-2
		Harris Hoop-truss
	Radial structure	Harris rigid ribs (TDRS and Galileo)
		Harris hinged ribs
		Tension-Truss (HALCA)
Peripheral ring and flexible radii	Georgian TU	
	NPO-EGS/Thales LDA	
Modular architecture	ETS-VIII	
	OKB-MEI Travers	
Membrane	Tensioned membrane	CRTS ESA/Cambridge
	Inflatable	Contraves, L'Garde
Deformable shell	Spring-back & variants	Hughes
		Cambridge University
Solid surfaces	Rotating petals or sectors	DAISY, MEA, Thin-shell
	Foldable tips	Selenia ASTP 20/30 GHz
	Longitudinal solid sections	XM radio
Hybrid surface	Solid-inflatable	PLC Dover
	Solid-mesh	Venera

Fig. 9.33 A selection of metal mesh reflectors. **a** AstroMesh reflector (Thuraya 12.25 m), Northrop-Grumman Space Technology. **b** Georgian reflector deployed on MIR 1999 (GTU 6 m). **c** ESA 12 m LDA. Thales Alenia/NPO-EGS. **d** Harris Hoop-truss reflector



latter technology, known as CFRS, leads to the shell-membrane concept.

The solid surfaces class contains several types of architecture, most of them based on articulated petals, folding tips, or separated parts that are rotated or translated in order to recreate a continuous solid CFRP surface. The solid reflecting surface portions are typically built in CFRP, thus allowing for high-frequency operation, often in the Ka band. The solid portion can also be taken as a central reflector of a hybrid construction, complemented by edge mounting of an inflatable annulus or deployable radial ribs that support a metal mesh.

9.15.4.3 Technologies

The technologies that enable large deployable reflectors (in addition to the actuators) are mainly those related to the manufacturing and integration of the reflecting surface, mitigation of shape imperfections, and improved stability of the supporting structures

- Metal knitted mesh materials and processes, production, integration, and tensioning
- Alternative reflecting surface technologies, such as the CFRS shell-membrane, which does not require tensioning
- Multi-feed illumination systems to perform digital beam reconfiguration in order to correct performance losses caused by disturbances (attitude, orbital and thermal)
- High-performance structural materials providing higher stiffness/mass and thermal conductivity by the use of pitch carbon fibers and carbon nano species.

Technologies maturing in other applications may be employed in large reflectors

- Inflatable components for very large apertures (above 20 m), considering the mass efficiency of inflatable elements
- Shape memory alloys
- Technologies for mechanical shaping in orbit.

Another aspect is the conceptual design and selection of the most appropriate architecture. This is not a technological point *per se*, but is very closely related to the selection of the reflecting surface and shape control. The truss-tension concept mounted on a peripheral ring or radial booms, seems to show a rather robust behavior that performs well compared to other architectures. The different methods for folding rings or booms have a strong impact on the mass, stiffness, shape stability, and deployment reliability. In Fig. 9.34, some of the commonly used folding schemes of articulated struts are summarized, including the pantograph and its variants. These folding schemes can be employed either in the radial booms or in the peripheral rings. Examples of their use can be identified in the existing reflectors listed in Table 9.5.

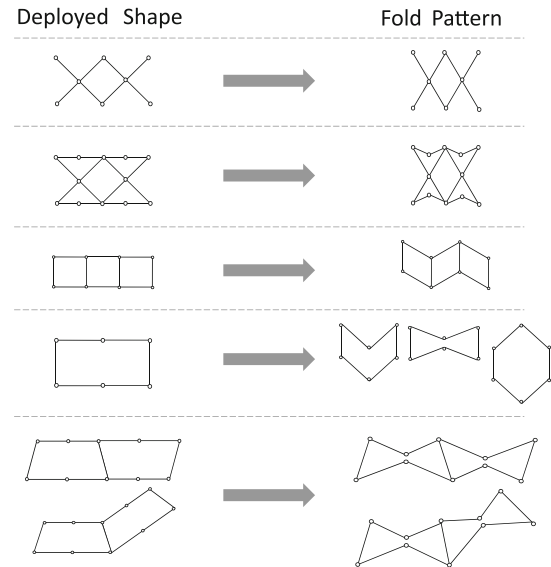


Fig. 9.34 Folding schemes of deployable linear elements

Inflatable structures are deployable structures whose deployment concept is based on inflation by gas. Inflatable space structures have been under development and evaluation for 50 years. Indeed their potential for low-cost flight hardware, high mechanical packaging efficiency, and low mass makes them very attractive. This was especially important in the context of launch vehicle capabilities in the early 1960s (limited volume and mass). In Europe, ESA showed an interest in large inflatable structures as early as the 1970s. An initial study performed by Contraves concluded “*the balloon technology introduces a novel approach to the manufacture of curved surface, an approach which is not restricted to the realization of antennae*”. Further work was then conducted and the feasibility of inflatable, space-rigidized structures (ISRS) was positively assessed and a reference object (a 10 m antenna reflector) was designed and subjected to a preliminary analysis. Materials investigations were performed to define cure catalysts, cure cycle, mechanical properties, and electrical properties. Finally, a ¼-scaled model (LOAD-3 for Large Offset Antenna Demonstrator 3 m diameter) was manufactured, deployed, and mechanical and electrical tests performed. Based on the very promising results, a full-scale model of a 10 m aperture antenna reflector was built. The final test to be performed, the flight experiment, was never conducted, but an in-orbit cure experiment was done. In addition to the development activities performed for the LOAD models, several projects have expressed interest in inflatable structures.

It is also of note that the in-orbit modules deployed by Bigelow Aerospace, intended to be fully crew-rated, use inflatable technology licensed, originally, from NASA in

the late 1990s but further developed with proprietary extensions by Bigelow Aerospace.

9.15.4.4 Design and Verification

The verification of inflatable structures is a challenge. Indeed, inflatable structures have properties that make their testing particularly difficult. These are pressurized systems, sometimes of very large dimensions, low mass, and with a change of state in case of rigidization. The effect of Earth gravity must be properly accounted for. Indeed, deployment is strongly influenced by gravity, and in some cases these structures are not able to sustain their own weight (especially before rigidization). They therefore require gravity compensation systems. Atmospheric pressure also affects the test results, as it might be of several orders of magnitude larger than the internal pressure required to inflate a thin membrane in vacuum (a few Pa). Tests under vacuum may be impracticable for large structures.

The test set-up itself is also a challenge for inflatable structures. Indeed, the instrumentation (strain gauges, accelerometers, etc.) cannot be used because its presence would influence the mass distribution (the accelerometer mass might be of the same order of magnitude as the membrane mass) and the stiffness (through the cabling). The solution is to use contact-less measurement techniques. Depending on the rigidization technique, testing on the ground may be difficult. Some techniques like ultraviolet or thermal curing, or metal-layer stretching are irreversible. Techniques relying on the space environment for rigidization, e.g. solar ultraviolet or thermal curing and dehydration require a simulation of this environment for on-ground testing.

The current deployment methods are directed at members such as tubes and struts. Normally these members are used to move the remainder of the inflatable system into position for inflation. Deployment methods must keep the deploying inflatable structure within a predictable envelope, and provide a well-defined deployment rate that is slow enough to prevent significant loads on the spacecraft. They must also provide restraint for the structure during launch. Several techniques are available

- Roll-out method: This method is similar to the well-known party favor. Embedded mechanisms (springs, Velcro, etc.) cause the tube to unroll in a prescribed plane and provide resistance to unrolling.
- Mandrel method: The inflated beam is extended in a straight telescopic motion with some degree of beam stiffness during the deployment.
- Fan-folded method: The deployment resistance is provided by the bending strength of the tube itself.

- Accordion method: The boom is folded in an accordion-pattern. The deployment resistance is then guaranteed by releasing the accordion petals in a controlled manner.

9.16 Further Reading

9.16.1 Standards

- Structural General Requirements ECSS-E-ST-32C Rev. 1, 15/11/2008
- Fracture Control ECSS-E-ST-32-01C Rev. 1, 6/3/2009
- Structural Design and Verification of Pressurized Hardware, ECSS-E-ST-32-02C Rev. 1, 15/11/2008
- Structural Finite Element Models, ECSS-E-ST-32-03C, 31/7/2008
- Materials, ECSS-E-ST-32-08C Rev. 1, 21/7/2008
- Structural Factors of Safety for Space Flight Hardware, ECSS-E-ST-32-10C Rev. 1, 6/3/2009
- Modal Survey Assessment, ECSS-E-ST-32-11C Rev. 1, 21/7/2008,
- Space product assurance: Materials, mechanical parts and processes, ECSS-Q-ST-70C, March 2009
- NASA Standard Materials and Processes Requirements for Spacecraft, NASA-STD- (I)-6016
- Verification, ECSS-E-ST-10-02C, 6/3/2009
- Testing, ECSS-E-10-03A, 15/2/2002 (superseded)
- Testing, ECSS-E-ST-10-03C, Draft 12.5, 4/3/2011 (in review)

9.16.2 Handbooks

- Adhesive Bonding Handbook, ECSS-E-HB-32-21A, 20/3/2011
- Insert Design Handbook, ECSS-E-HB-32-22A, 20/3/2011
- Threaded Fasteners Handbook, ECSS-E-HB-32-23A, 10/4/2010
- Buckling of Structures, ECSS-E-HB-32-24A, 24/3/2010
- Spacecraft Load Analysis, ECSS-E-HB-32-26A, TBD issue date
- Structural Acoustics Design Manual, ESA PSS-03-204, March 1996
- Mechanical Shock Design and verification Handbook, ESA Contract No 20503/06/NL/Sfe, 15/9/2011 (will be issued as ECSS Handbook)
- Space Product Assurance, Data for selection of space materials and processes ECSS-Q-70-71A, Rev 1, 18 June 2004

- Space Product Assurance, Materials, mechanical parts and processes ECSS-Q-ST-70C, 6 March 2009
- Space Product Assurance, Material selection for controlling stress-corrosion cracking, ECSS-Q-ST-70-36C, 6 March 2009
- Space Product Assurance, Safety, ECSS-Q-ST-70-40C, 6 March 2009
- Space Product Assurance, Standard methods for mechanical testing of metallic materials ECSS-Q-ST-70-45C, Rev 1
- Composite Materials Handbook, MIL-HDBK-17-2F

9.16.3 NASA Handbooks and Papers

- NASA STD-7003 Pyro shock Test Criteria, 1999
- NASA SP-8019 Buckling of Thin-Walled Truncated Cones
- NASA SP-8007 Buckling of Thin-Walled Circular Cylinders
- H.N. Abramson, The Dynamic Behavior of Liquids in Moving Containers, with applications to Space Vehicle Technology, NASA SP-106 (1967).

9.16.4 Books

- Osgood, C.C (1966), Spacecraft Structures, Prentice-Hall.
- T.P. Sarafin, Spacecraft Structures and Mechanisms, From Concept to Launch, (1995), Space Technology Series, ISBN 0-7923-3476-0.
- Agrawal, B.N. (1986) Design of Geosynchronous Spacecraft, Prentice Hall, ISBN 0-13-200114-4.
- J.J. Wijker, Spacecraft Structures, (2008) Springer, ISBN 978-3-540-75552-4.
- R.D. Cook, D.S. Malkus, M.E. Plesha, Concepts and Applications of Finite Element Analysis, (1989), John Wiley, ISBN0-471-84788-7.
- Kwon, Y.W., Bang, H. The Finite Element method Using MATLAB, CRC Press, 2000, ISBN 0-8493-0096-7.
- Preumont, A. (2011) Vibration Control of Active Systems, Springer, ISBN 978-94-007-2032-9.
- T.H.G. Megson, Aircraft Structures for Engineering Students, third edition, (1999), Butterworth Heinemann, ISBN 0-340-70588-4.
- Gere, J.M., Timoshenko, S.P., Mechanics of materials, third edition, Chapman & Hall, 0-412-36880-3.
- Den Hartog, J.P., Strength of Materials, Dover, 1961, ISBN 0486607550.
- Blevins, R.D., Formulas for Natural Frequencies and Mode Shape, Krieger Publishing, 1995, ISBN 0-89464-894-2.

- Lyon, R. H., and DeJong, R. G., Theory and Application of Statistical Energy Analysis, 2nd ed., Butterworth-Heinemann, Boston, 1995.
- Wu, T.W. (2005) Boundary Element Acoustics, WIT press, ISBN 1-85312-570.
- Thomson, W.T., Dahleh, M.D., (1993) Theory of Vibrations with Applications, 5th edition, Prentice-Hall, ISBN 0-13-651068-X.

9.16.5 Papers from Journals and Conferences

- Kaplow, C.E. and Velman, J.R. (1980). Active local vibration isolation applied to a flexible space telescope. American Institute of Aeronautics.
- G. Ramusat, L. Innocenti, M. Caporicci, H. Krings (1999) An overview of the Agency Technology Development Programmes in Materials for Reusable Launch Vehicles, 18th European Conference on Materials for Aerospace Applications, June 1999.
- Sairajan, K.K., et al. (2005) Optimum Design of a Composite Base Structure of a Spacecraft, Altair CAE Users Conference 2005, August 11–13, Bangalore.
- ESA approach to the prevention of stress-corrosion-cracking in spacecraft hardware, G. Bussu, B.D. Dunn, 2002.
- Experimental assessment of the susceptibility to stress-corrosion-cracking of Ti-6Al-4V alloy exposed to MON-1 propellant tank environment—background and test design, G. Bussu, D. Stramaccioni, I. Kälsch, 2004.
- The degradation of metal surfaces by atomic oxygen, Proceedings of the Third European Symposium on Spacecraft Materials in Space Environment, Noordwijk, The Netherlands, 1985, A. De Rooij.
- Composite structures research and technology activities in ESA, A. Obst, L. Daniel, J. S. Prowald, G. Sinnema, International Astronautical Congress (IAC), 2003.

References

1. Blevins, R.D., Formulas for Natural Frequencies and Mode Shape, Krieger Publishing, 1995, ISBN 0-89464-894-2
2. Blevins, R.D., Formulas for Natural Frequencies and Mode Shape, Krieger Publishing, 1995, ISBN 0-89464-894-2
3. Blevins, R.D., Formulas for Natural Frequencies and Mode Shape, Krieger Publishing, 1995, ISBN 0-89464-894-2
4. Blevins, R.D., Formulas for Natural Frequencies and Mode Shape, Krieger Publishing, 1995, ISBN 0-89464-894-2
5. Den Hartog, J.P., Strength of Materials, Dover, 1961, ISBN 0486607550
6. Gere, J.M., Timoshenko, S.P., Mechanics of materials, third edition, Chapman & Hall, 0-412-36880-3

7. T.H.G. Megson, *Aircraft Structures for Engineering Students*, third edition, (1999), Butterworth Heinemann, ISBN 0-340-70588-4
8. R.D. Cook, D.S. Malkus, M.E. Plesha, *Concepts and Applications of Finite Element Analysis*, (1989), John Wiley, ISBN0-471-84788-7
9. Kwon, Y.W., Bang, H. *The Finite Element method Using MATLAB*, CRC Press, 2000, ISBN 0-8493-0096-7
10. Wu, T.W. (2005) *Boundary Element Acoustics*, WITpress, ISBN 1-85312-570
11. Lyon, R. H., and DeJong, R. G., *Theory and Application of Statistical Energy Analysis*, 2nd ed., Butterworth-Heinemann, Boston, 1995.
12. Thomson, W.T., Dahleh, M.D., (1993) *Theory of Vibrations with Applications*, 5th edition, Prentice-Hall, ISBN 0-13-651068-X
13. Kaplow, C.E. and Velman, J.R. (1980). Active local vibration isolation applied to a flexible space telescope. American Institute of Aeronautics.
14. Preumont, A. (2011) *Vibration Control of Active Systems*, Springer, ISBN 978-94-007-2032-9
15. ESA approach to the prevention of stress-corrosion-cracking in spacecraft hardware, G. Bussu, B.D. Dunn, 2002
16. Experimental assessment of the susceptibility to stress-corrosion-cracking of Ti-6Al-4 V alloy exposed to MON-1 propellant tank environment – background and test design, G. Bussu, D. Stramaccioni, I. Kälsch, 2004
17. The degradation of metal surfaces by atomic oxygen, Proceedings of the Third European Symposium on Spacecraft Materials in Space Environment, Noordwijk, The Netherlands, 1985, A. De Rooij
18. *Composite Materials Handbook*, MIL-HDBK-17-2F
19. Composite structures research and technology activities in ESA, A. Obst, L. Daniel, J. S. Prowald, G. Sinnema, International Astronautical Congress (IAC), 2003

Mukund R. Patel

10.1 Power System Basics

The electrical power system (EPS) generates, stores, conditions, controls, and distributes power within the specified voltage band to all bus and payload equipment. The protection of the power system components in case of all credible faults is also included. The basic components of the most widely used power system are (1) solar array, (2) solar array drive, (3) battery, (4) battery charge and discharge regulators, (5) bus voltage regulator, and (6) switches, fuses, and distribution harness.

Power requirements in very early satellites were several watts. In today's communication satellites, it can be 20 kW or higher. Some strategic defense spacecraft power requirement may be in hundreds of kilowatts and some defense concepts require hundreds of megawatts of burst power. Solar radiation is the only external source of primary energy available in space with reasonable flight heritage, although other concepts such as conductive wires have been considered. Any spacecraft not using solar energy must carry on-board its own source of energy, such as the primary battery, radioisotope, nuclear reactor, or chemical fuel. The conversion of the primary energy into electrical energy may be photovoltaic (PV), thermoelectric (TE), dynamic alternator, or thermionic. Some energy storage may also be required in many spacecraft to meet the load power requirement during eclipse or during any peak demand period. Energy storage has been primarily by electrochemical battery, although regenerative fuel cell and flywheel technologies are under development.

From available options that are compatible with a given mission and its environment, the satellite-level optimization study determines the best combination of primary energy

source, energy conversion, and energy storage technologies. Final selection must meet multiple criteria, but the primary criteria are always low mass and low life cycle cost. Such selection is largely influenced by the power level \times mission duration product as shown in Fig. 10.1, where the dividing lines among various options have large overlaps. Although the PV-battery power system is the most common for Earth-orbiting satellites, a variety of alternative power system technologies have been developed and flown for various space missions [1]. The practical limit and performance characteristics of major power system options are summarized in Table 10.1.

The U.S. Department of Defense and NASA have also funded the development and testing of proto-flight solar array designs that could yield a specific power of over 100 W/kg, a factor of 3 greater than the state of the art and a factor of 5 greater than the state of the practice. The advanced designs under consideration integrate three promising technologies [2]: (1) flexible copper indium diselenide thin-film PV cells, (2) smart mechanisms using shape memory metal, and (3) multi-functional lightweight structures. An important criterion in application of a new technology is the flight qualification status. Any new component is subject to time-consuming and expensive testing to prove its ability to withstand launch and space environments.

10.2 Photovoltaic-Battery System

The solar energy and PV-battery power system is widely used in Earth orbiting satellites. The solar flux received from the Sun varies with the distance squared. Earth's orbit around the Sun is approximately circular with a slight eccentricity of 0.01672. The distance, therefore, varies within ± 0.01672 times the average distance between the Sun and the Earth, which is 149.6 million km, defined as one astronomical unit (au) of distance. Thus, the solar flux varies

M. R. Patel (✉)
U.S. Merchant Marine Academy, Kings Point, NY, USA

Fig. 10.1 Optimum primary energy sources for various power levels and mission durations

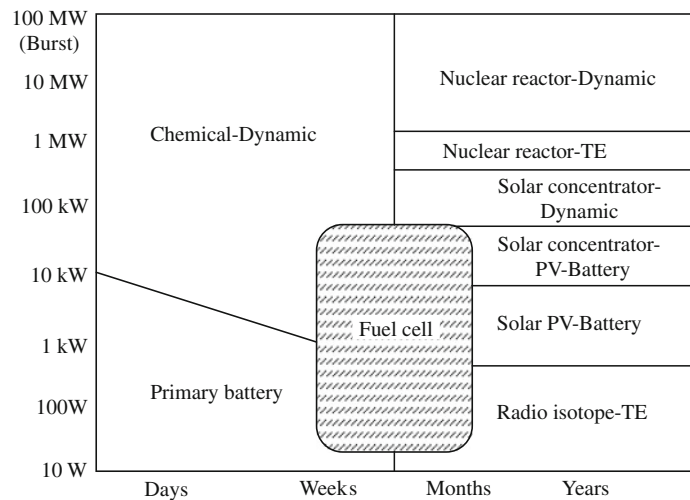


Table 10.1 Practical limit and performance of various power system options

Power system option	Practical power limit (kW)	Net system efficiency (%)	Specific power
Solar-PV	20–30	15–30	5–10 W/kg
Isotope-TE	1	7–15	5–10 W/kg
Nuclear-TE	100–300	7–15	–

over $(1 \pm 0.01672)^2$ or 1 ± 0.034 of the yearly average. For many years, the average solar radiation in Earth orbit was taken as $1,358 \pm 5 \text{ W/m}^2$ on a surface normal to the Sun. Measurements reported by Frohlich [3] showed a higher average value of $1,377 \pm 5 \text{ W/m}^2$, however the conservative number of $(1,358 - 5) = 1,353 \text{ W/m}^2$ continues in wide use, particularly for power system sizing. Recall from Sect. 3.2.3 that the ASTM E490 Standard Solar Constant and Zero Air Mass (AM0) Solar Spectral Irradiance has an integrated power of $1,366.1 \text{ W/m}^2$; an ISO standard is also available, see ISO-21348. Since the ecliptic and equatorial planes are inclined to each other by 23.45° , the angle of incidence of the sunlight on an uncontrolled geostationary satellite’s solar arrays varies from 66.55° to 90° . The corresponding incident solar flux varies from 91.75 % on a solstice day to 100 % on an equinox day. However, the satellite on equinox days encounters the longest eclipse once per day when the Earth blocks the sunlight from illuminating the satellite.

Alternative PV-battery architectures are depicted in Fig. 10.2. The direct energy transfer (DET) from the solar array to the load is best for the overall system efficiency. In the DET category are (1) a sunlight regulated bus in which the excess solar array current (if any during sunlight) is shunted to the ground in order to maintain the bus voltage in a narrowband and the bus voltage during eclipse is the

battery voltage. And, (2) a fully regulated bus in which the bus voltage is regulated within a narrowband during the entire orbit. The other option is (3) a peak power tracking (PPT) architecture in which a series regulator matches the load and the source characteristics to extract the maximum power from the solar array. Power loss of several percent in the PPT regulator sometimes nullifies the gain in power from solar array. The PPT architecture may be suitable for a mission operating over wide orbit parameters. However, the sunlight or fully regulated bus generally leads to the optimized design for large high-power satellites.

10.2.1 Solar Array

The solar array is made of numerous PV cells mounted on a base substrate and connected in a series-parallel combination to obtain the desired voltage and current. Each cell can be of any of the following types: (1) single crystal silicon, (2) gallium arsenide, (3) semi-crystalline or polycrystalline, (4) thin film, (5) amorphous, or (6) multi-junction. As of 2012, a new class of ultra-light, high-efficiency solar cell has been developed by the U.S. Department of Energy. It is the inverted metamorphic multijunction (IMM) with the conversion efficiency of $42.3 \pm 2.5 \%$ as measured by the U.S. National Renewable Energy Laboratory. The tests were made on *In-GaP/GaAs/InGaAs* three-junction cells with concentration of 406 Suns on Earth (atmospheric mass, AM, 1.5) and cell temperature of 250°C [4]. These cells consist of multiple thin films in layers that allow the cell to capture more of the solar spectrum and convert it into electrical power. The maximum reported efficiency is slightly better than 41.4 % achieved by the Fraunhofer Institute for Solar Energy Systems.

The IMM cells, primarily developed for terrestrial applications at present, are also developed for the space satellite market by companies like Emcore and Spectrolab. The IMM solar array in space without concentration could

Fig. 10.2 Alternative architectures for spacecraft power. **a** DET-sunlight regulated bus. **b** DET-fully regulated bus. **c** Peak power tracking bus

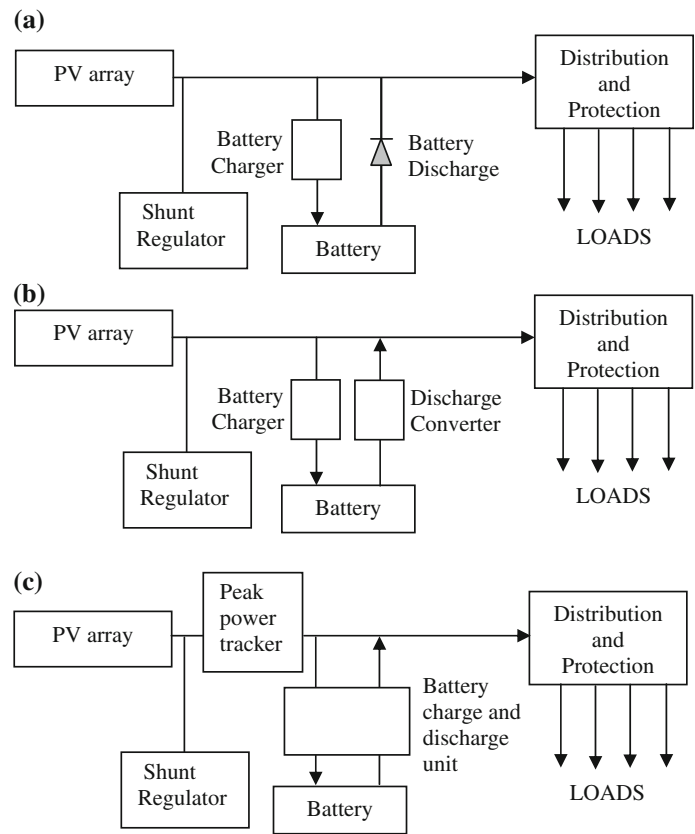
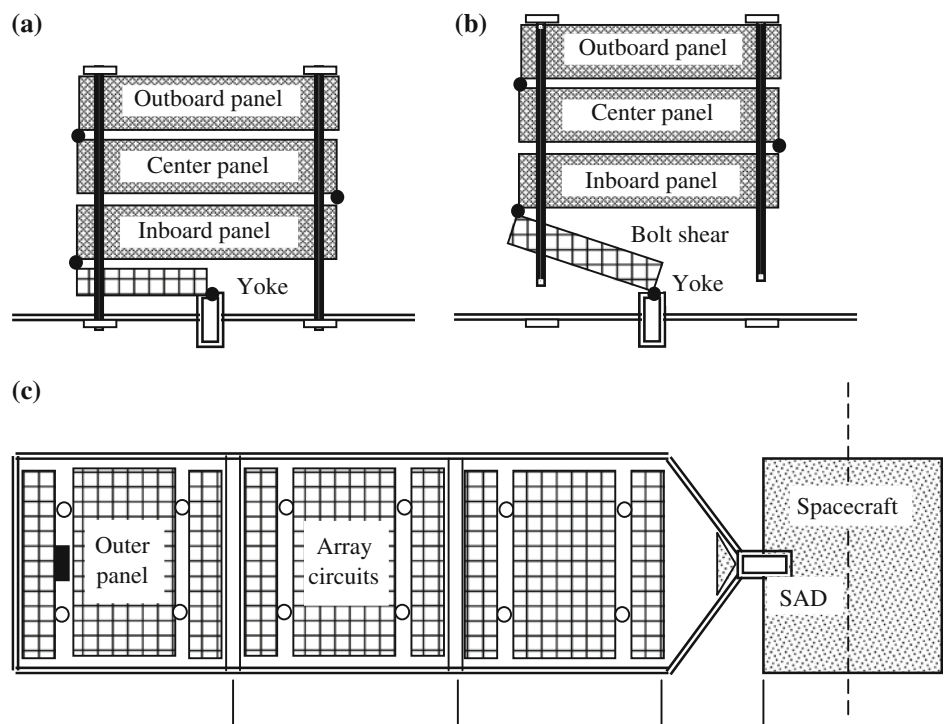


Fig. 10.3 Flat panel solar array wing: **a** stowed (array is held in place by four hold-down release devices during launch and transfer orbit), **b** being deployed (four hold-down release devices are actuated when final orbit is achieved and yoke begins to unfold), and **c** fully deployed (after center panel is fully outfolded and locked, the outboard panel fully unfolds and locks. Hinge position microswitches send the array-deployed signal)



be 33–35 % efficient and be incorporated into a satellite’s skin or unfurl like an awning. This would eliminate the need for conventional wing-shaped solar arrays with heavy metal frames and associated mechanisms.

Solar array construction can be (1) rigid flat panels, (2) body mounted panels, (3) flexible blanket type array, (4) inflatable balloon type array, or (5) concentration array. Most arrays are made with crystalline silicon or gallium

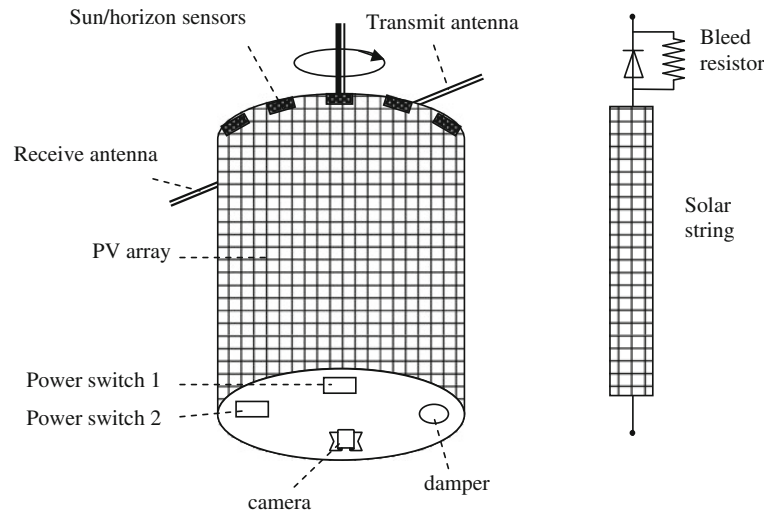


Fig. 10.4 Body mounted solar array with bleed resistor in each solar string

Table 10.2 Key features of 3-axis stabilized and spin-stabilized satellites

3-axis stabilized flat wings	Spin-stabilized round body
Bias or zero momentum maintains the stability	Inherently stiff due to rotational inertia
Complex attitude control	Simple mechanical structure
Full solar array normal to the Sun generates power all the time	Only 1/3rd of the solar array generates power at any time
Can have high power by adding solar panels on wings	Power limited by body size that fits the launch vehicle
Great flexibility in design	Less flexibility in design
Suitable for large satellites	Suitable for small satellites

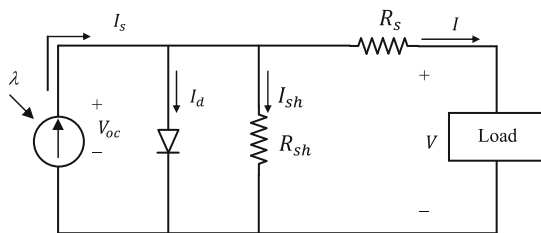


Fig. 10.5 Equivalent electrical circuit of solar cell

arsenide cells on rigid panels, where the solar panels can be flat wings (Fig. 10.3) in 3-axis stabilized spacecraft, or mounted on a round body surface (Fig. 10.4) in spin-stabilized spacecraft. The key features of these two basic types of solar panels are given in Table 10.2.

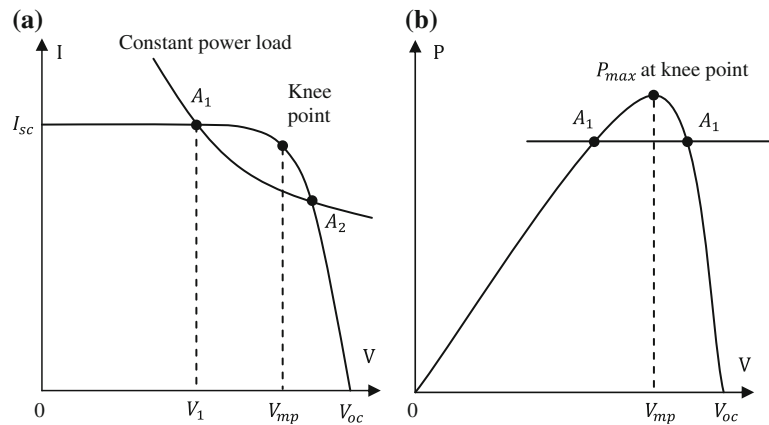
The steady state equivalent electrical circuit of a PV cell (and also of a PV panel by series–parallel scaling) is shown in Fig. 10.5. The cell acts as a constant current source shunted by a diode. In the circuit parameters, the series resistance R_s represents the internal resistance to the current flow, which is primarily due to the resistivity of the material. The shunt resistance represents the leakage current across the junction. It depends on the p-n junction depth, the

impurities, and the contact resistance. The value of R_{sh} is inversely related with the leakage current to ground. In an ideal PV cell, $R_s = 0$ (no series loss), and $R_{sh} = \infty$ (no leakage to ground). In a typical high quality 2.0×2.5 cm silicon cell, $R_s = 0.05\text{--}0.10 \Omega$ and $R_{sh} = 200\text{--}300 \Omega$. The PV conversion efficiency is sensitive to small variations in R_s , but is insensitive to variations in R_{sh} . The value of R_{sh} affects the constant current slope, whereas the value of R_s affects the constant voltage slope. A small increase in R_s can decrease the PV output significantly. Since the magnitudes of the diode current and the resistances R_s and R_{sh} vary with temperature, the cell output and the conversion efficiency decrease with increasing temperature.

10.2.1.1 P–V and I–V Characteristics

The current versus voltage (I–V) characteristic of the PV cell in sunlight is shown in Fig. 10.6a, where two important parameters for characterizing the cell performance are the open circuit voltage V_{oc} and the short circuit current I_{sc} . The short circuit current is measured by shorting the output terminals and measuring the terminal current under full illumination. Ignoring the small diode and ground leakage currents under zero terminal voltage, I_{sc} is the photocurrent

Fig. 10.6 Solar array output characteristics and constant power load curves. **a** I–V characteristic. **b** P–V characteristic



I_s . The current under this condition is the maximum current the cell can deliver. The bottom right of the curve at zero current is the open circuit voltage measured with the output terminals open. The maximum photo voltage is produced under the open circuit voltage. The I–V curves over a full range are developed from the test data at various illuminations, temperatures, and ionized radiation doses.

The product of voltage and current outputs is the output power, which is plotted in Fig. 10.6b. The cell produces no power at zero voltage or zero current, and produces the maximum power at a voltage corresponding to the knee point of the I–V curve. This is why the PV power circuits are designed such that the panel operates closed to the knee-point, slightly on the left-hand side, where the cell operates approximately as a constant current source.

The basic requirements of solar cells for space applications are generally described in MIL-STD-83576 by the U.S. Air Force. The cell specifications for general use in space are defined by AIAA Standards-115 and -116. The ISO Technical Committee C20 has issued ISO-15387 for the aircraft and space vehicles and this is endorsed by American National Standard Institute. NASA's Jet Propulsion laboratory (JPL) often performs such tests in aircraft and balloons at 35 km height. Comparing such tests conducted by one organization with another is difficult. For example, in a round-robin test by six organizations to ISO-15387, including tests conducted at JPL and in the PV engineering test bed at NASA's Glenn Research Center, the results showed up to 3 % deviations on the same cell. The cell temperature coefficient is extremely sensitive to solar spectrum, and can vary several fold, making the spectrum duplication in tests extremely important. For this reason, power engineers allow 3 % margin for standard cells, and perhaps more for multi-junction cells, as the possible error in power generation estimates using the test results.

10.2.1.2 Array Performance

Major factors influencing the electrical performance of the solar array are solar intensity, Sun angle, and the operating temperature. The I–V characteristic of a PV array reduces in

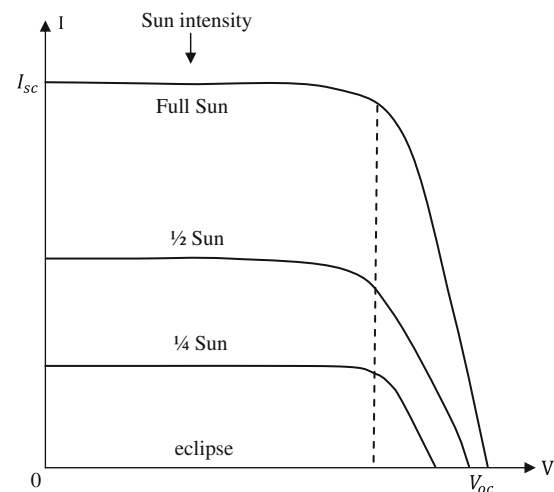


Fig. 10.7 Solar array current versus voltage at various illumination levels

magnitude at lower Sun intensity with a small reduction in voltage as shown in Fig. 10.7. However, the photo conversion efficiency of the cell is insensitive to solar illumination in the practical working range. Figure 10.8 shows that the efficiency is practically the same at full Sun ($1,353 \text{ W/m}^2$) and 1/2 Sun, and starts to fall off rapidly only below 1/4 Sun (340 W/m^2).

The cell output current is given by $I_s = I_0 \cos \theta$, where I_0 is the photocurrent with normal Sun ($\theta = 0$). The cosine law holds well for Sun angles ranging from 0 to about 50° , beyond which the electrical output deviates significantly from the cosine value. The cell generates no power beyond 85° , although the mathematical prediction would give 7.5 % power generation. The actual power versus angle curve is called the Kelly cosine, which is useful to assess accurately the power available from the Sun at low angles during transfer orbit.

With increasing temperature, the short circuit current of the cell increases, whereas the open circuit voltage decreases as shown in Fig. 10.9; as the increase in current is much less than the decrease in voltage, the net effect is the decrease in

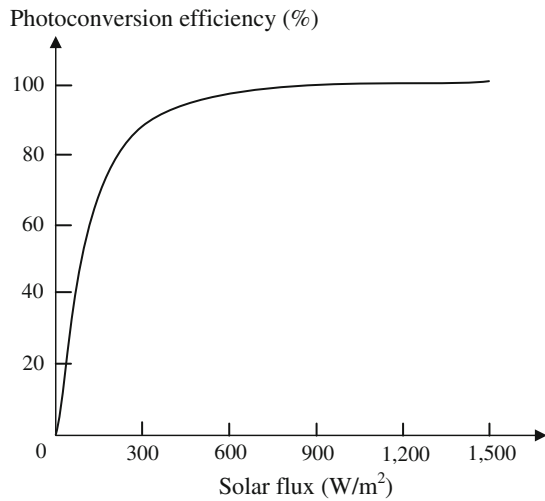


Fig. 10.8 Photoconversion efficiency remains constant over wide range of solar flux up to 1/4 Sun

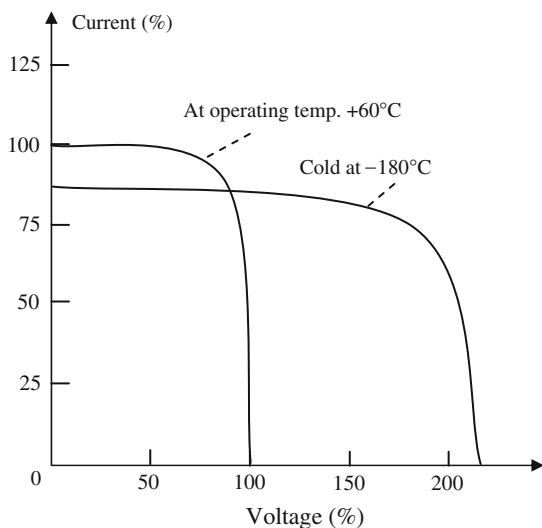


Fig. 10.9 Temperature effect on I-V curve of solar array

power, which is quantitatively evaluated by examining the effects on current and voltage separately. Say that I_0 and V_0 are the short circuit current and the open circuit voltage at reference temperature T , and α and β are their temperature coefficients in units of $A/^\circ C$ and $V/^\circ C$, respectively. If the operating temperature is increased by ΔT , then the new current and voltage are given by $I_{sc} = (I_0 + \alpha \Delta T)$ and $V_{oc} = (V_0 + \beta \Delta T)$. Since the operating current and voltage change approximately in the same proportion as the short circuit current and open circuit voltage, respectively, the new power $P = VI = (I_0 + \alpha \Delta T) \cdot (V_0 + \beta \Delta T)$. Ignoring the small term containing the product of α and β , $P = V_0 I_0 + \alpha \Delta T V_0 - \beta \Delta T I_0$, which reduces to a simple form

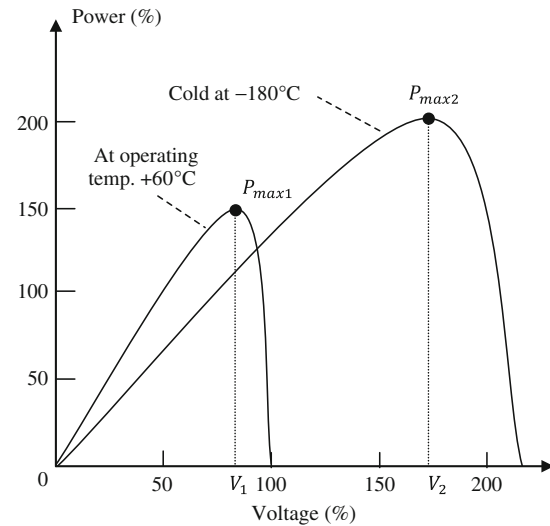


Fig. 10.10 Temperature effects on P-V curve of solar array

$$P = P_0 - [(\alpha V_0 - \beta I_0) \Delta T] r^n = 0.4 + 0.3(2^n). \quad (10.1)$$

For a typical 2×4 cm single-crystal silicon cell, α is $250 \mu A/^\circ C$ and β is $2.25 mV/^\circ C$. Therefore, the power varies approximately as $P = P_0(1 - 0.005 \Delta T)$, which indicates that for every $1^\circ C$ rise in the operating temperature, the silicon cell power output decreases by about 0.50 %.

Figure 10.10 depicts the power output versus voltage characteristic at two operating temperatures. It shows that the maximum power generated at the lower temperature is higher. Thus, cold temperatures are better for the PV cell for power generation. However, the two P_{max} points are not at the same voltage. In order to extract maximum power at all temperatures, the PV system must be designed such that the array output voltage can increase to V_2 for capturing P_{max2} at lower temperature, and can decrease to V_1 for capturing P_{max1} at higher temperatures. If the array is operating at a fixed regulated voltage, the higher power generation capability at cold temperatures cannot be utilized by the loads, and the excess power from the cell must be wasted in shunt circuits. The peak power all the time, regardless of the temperature, can be captured and utilized only by using the PPT architecture.

The array undergoes a wide temperature cycle in each orbit. During sunlight, the front face rises to $50\text{--}60^\circ C$ and the back face to $40\text{--}50^\circ C$. The solar array temperature is determined by the thermal equation: (Solar flux + Earth's albedo + Earth's thermal radiation + Heat coming from adjacent components of the spacecraft) = (Electrical power output + Heat radiated back into space). During eclipse in GEO, the temperature drops exponentially to as low as $-175^\circ C$. The time constant depends on the mass composition of the array components, and is typically 30–60 min. The front to back face temperature gradient for a rigid array with face sheets made of either aluminum (obsolete) or

Table 10.3 Degradation factors for 900 W satellite with a 15 year operational life in medium Earth orbit; 10,900 nm, 69° inclination; GPS II F orbit

I _{sc} factors	
<i>Natural radiation</i> ^a	
Assembly mismatch loss ^b	0.98
Cover glass charge particles	0.99
Cover glass coating (ITO) ^b	0.98
Ultraviolet rays	0.97
Propellant contamination	0.98
Micrometeoroid damage	0.98
V _{oc} factors	
<i>Natural radiation</i> ^a	
Cover glass charge particles	0.99
P _{max} factors	
<i>Natural radiation</i> ^c	
Wiring loss (cell-to-cell) ^b	0.98

^a Determined from the radiation fluence over the mission life

^b Beginning of life

^c Approximate value equal to the product of the I_{sc} and V_{oc} degradation factors can be used in top-level calculations only. In detailed calculations, it is accounted through the I_{sc} and V_{oc} factors

graphite epoxy (new designs) can be 5–10 °C under steady Sun and up to 20 °C on sunlight snap after eclipse. Various techniques are used to control the temperature of spacecraft parts over a full orbit. Figure 10.9 indicates that a typical cold array facing the Sun immediately after coming out of a long eclipse in GEO develops about twice its normal operating voltage in steady sunlight.

Solar array performance degrades due to two distinct groups of factors. In the first group are the initial degradations at beginning-of-life (BOL) due to (1) cell mismatch in the assembly, (2) cell-to-cell wiring loss, (3) power loss in array and boom wires, and (4) plasma effects that cause leakage current from the array to space. In the second group are the accumulated degradations up to the end-of-life (EOL) due to various environmental effects, such as (1) cumulative radiation dose of the ionized particles, (2) effect of ultraviolet rays on the optical properties of cover glass, (3) mechanical stress cycles causing soldered joints to crack over time, (4) impacts of micrometeoroids and debris damaging solder joints and reducing power generating area of the cells, (5) flue gases changing the optical properties of cover glass, (6) bypass diode failure causing loss in the string current. Table 10.3 gives representative values of major degradation factors for a typical medium Earth orbit, from which the total degradation can be obtained as

$$\begin{aligned} \text{Total degradation} &= \text{BOL degradation} \\ &+ \text{In-service degradation to EOL.} \end{aligned} \quad (10.2)$$

10.2.1.3 Peak Power Tracking

Sun tracking is required in order for the solar array to face the Sun continuously as the spacecraft orbits the Earth. This is done by an actuator that follows the Sun like a sunflower. There are two types of Sun trackers: (1) single-axis gimbals, which follow the Sun from east to west during the day, and (2) dual-axis gimbals that track the Sun from east to west during the day and from north to south during the seasons of the year. The dual axis tracking is done by two linear actuator motors, which aim at the Sun within 1° of accuracy. The EPS provides a means of independently orienting and rotating the deployed north and south solar wings about the pitch axis of the spacecraft. After the deployment, ground control is involved in aligning the solar array to the Sun. Once aligned, the array rotates by the clock time without ground intervention to maintain the Sun orientation. Some Sun pointing error cannot be avoided even after acquiring the normal Sun by the Sun sensor and then tracking by the solar array drive. The error generally comes from the cell flatness error of about 2° and the gimbal tolerance error of about 3°. The total 5° error must be accounted for in the array design. The cosine of 5° is 0.996, which means a 5° Sun pointing error will reduce the power generation by 0.4 %.

The gimbal's motor drives the array module to face the Sun to collect the maximum solar flux. However, that alone does not guarantee the maximum power output from the module. The module must electrically operate at the voltage that corresponds to the peak power point P_{max} under the given operating conditions. If the array is operating at voltage V and current I on the I–V curve, the power generation is $P = VI$ watts. If the operation moves away from the above point, such that the current is now $(I + \Delta I)$, and the voltage is $(V + \Delta V)$, the new power is $P + \Delta P = (V + \Delta V)(I + \Delta I)$. After ignoring a small term, this equation simplifies to $\Delta P = (\Delta VI + \Delta IV)$. The ΔP should be zero at the peak power point, which necessarily lies on a locally flat neighborhood as shown in Fig. 10.6b. Therefore

$$\text{At } P_{max} \text{ point, } \frac{dP}{dV} = 0, \text{ which reduces to } \frac{dV}{dI} = -\frac{V}{I}. \quad (10.3)$$

Note that (dV/dI) is the dynamic impedance and (V/I) is the static impedance of the PV array, and the P_{max} point is at the knee-point shown in Fig. 10.6a.

10.2.2 Battery

Energy storage is required in order to meet the spacecraft load demand during the launch/injection phase, during

eclipses, and when the demand exceeds the power generation at any time. The most widely used energy storage technology is the rechargeable battery that stores energy in electrochemical form. The battery is made of numerous electrochemical cells assembled in series–parallel combination to obtain the required voltage and current. The cell stores energy at a low electrical potential. The cell voltage depends solely on the electrochemistry, and not on the physical size. Commonly used electrochemistries produce 1.5–4.2 V when fully charged. The cell's ampere-hour (Ah) storage capacity, denoted by C , depends on the physical size. It is defined as the Ah charge that the cell can deliver at room temperature until it reaches a cut-off voltage of about two-thirds of the fully charged cell voltage. The battery can deliver C amperes for 1 h or C/n amperes for n hours. The cell capacity measures the Ah output at the terminals, not what is stored between the plates. A 1.5 V cell discharged to 1.0 V delivers practically the full capacity of the cell, and delivers only a few percent more if drained further to 0.1 V.

The battery voltage rating is stated in terms of the average voltage during discharge. A higher-voltage battery requires a greater number of cells in series. The product of voltage and the Ah rating gives the energy rating in watt-hours (Wh) that the battery can deliver to a load from the fully charged state. The battery charge and discharge rates are stated in fractions of the capacity. For example, charging a 100 Ah battery at a 10 A rate is said to be charging at $C/10$ rate. Discharging this battery at $C/2$ rate means drawing 50 A. At this rate, the battery will be fully discharged in 2 h. The battery depth of discharge $DoD = (1 - SoC)$, where

$$SoC = \frac{\text{Battery state of charge}}{\text{Ah capacity remaining in the battery}} = \frac{\text{Rated Ah capacity}}{\text{Rated Ah capacity}} \quad (10.4)$$

Major rechargeable batteries used in the spacecraft industry at present are (1) nickel–cadmium (NiCd), (2) nickel–hydrogen (NiH₂), and (3) lithium-ion (Li-ion). New electrochemistries are continuously researched for space applications [5], and for a variety of ground-based applications—consumer electronics, electric vehicles, utility load leveling, and renewable power systems. Lithium-polymer (Li-poly) and nickel-metal-hydride (NiMH) are two such examples in the commercial world.

The following figures of merit are often used in comparing the relative performance of various electrochemistries: (1) specific energy or gravimetric energy density = energy stored per unit mass, Wh/kg, (2) energy density or volumetric energy density = energy stored per unit volume, Wh/L, (3) cycle life = number of charge/discharge cycles the battery can deliver while maintaining

the minimum required voltage, and (4) specific power and power density = power the battery can practically deliver per kilogram of mass and liter of volume, respectively. It is sometimes necessary to think in terms of the power parameters also, since the internal resistances of the battery may limit the rate at which the energy can be discharged within practical design limits.

The NiCd battery has served as the workhorse of the spacecraft industry since the earliest missions, and is still used in some missions. However, since the mid-1980s it has been replaced by NiH₂ in general use due to NiCd's memory effect—loss of capacity after repeated use at low DoD. The NiH₂ provides deeper DoD for comparable cycle life, thus requiring lower Ah capacity, which translates into lighter weight. Today, the industry appears to be moving towards lithium based batteries for potentially 2–5 times the specific energy compared to NiH₂. Lithium-ion has a charge/discharge ratio and a round trip energy efficiency close to unity at low depth of discharge. However, no single electrochemistry can meet the wide range of space mission requirements. All chemistries will perhaps continue in use where they fit the best for a minimum mass and cost design.

10.2.2.1 Battery Performance

The battery works as a voltage source with small internal resistance. Its electrical circuit model has the internal electrochemical voltage E_i with internal resistance R_i in series. The E_i decreases and R_i increases linearly with the Ah discharge. Quantitatively

$$E_i = E_o - K_1 \cdot DoD \quad \text{and} \quad R_i = R_o + K_2 \cdot DoD \quad (10.5)$$

where E_o and R_o are internal voltage and internal resistance in a fully charged battery with $DoD = 0$, and K_1 and K_2 are electrochemistry constants to be found by curve-fitting the test data. The terminal voltage drops with increasing load current I , such that $V_{Terminal} = E_i - IR_i$, where E_i and R_i are functions of DoD. Thus, the terminal voltage is also a function of DoD as shown in Fig. 10.11 for NiH₂ cell in one full LEO orbit, discharging to various DoD levels during a 36-min eclipse and then fully charging up before the next orbit.

The design and operation of a battery requires certain safety considerations. The most important is not to overcharge the battery. Any overcharge above the trickle charge rate is converted into heat, which can explode the battery if allowed to build up beyond limit. This is particularly critical when the battery is charged directly from a dedicated photovoltaic module without a charge regulator in small science missions with short duration or infrequent eclipses. In such cases, the array rating is kept below the continuous trickle charge current that can be tolerated by the battery.

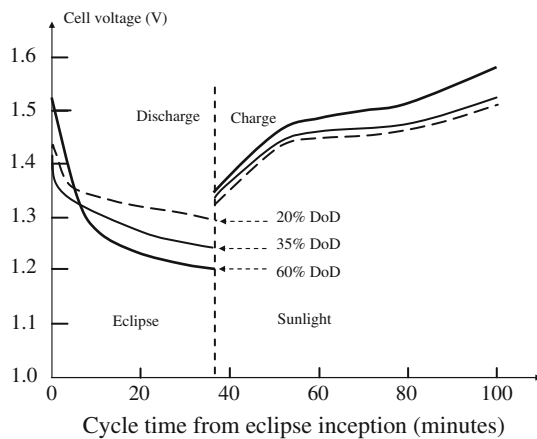


Fig. 10.11 NiH₂ cell voltage during one charge/discharge cycle at various DoD in LEO

10.2.2.2 Battery Life

The battery's primary mode of failure is associated with electrode wear due to repeated charge/discharge cycles. The cycle life depends strongly on the electrochemistry, depth of discharge, and temperature, as depicted in Fig. 10.12 for NiCd and NiH₂ batteries. The battery life also depends to a lesser degree on the electrolyte concentration, electrode porosity, and charge and discharge rates. The electrolyte concentration makes significant difference in the cycle life of NiH₂ cell. The cell with 26 % concentration gives a greater cycle life than one with 31 % concentration.

The number of charge/discharge cycles in a satellite equals the number of eclipses during the mission life. It is at least an order of magnitude greater in LEO than in GEO. Such long cycle life requirement in LEO can be achieved only by limiting the battery design to a low DoD, typically 30 % in LEO as compared to 80 % in GEO. For this reason, a LEO battery is proportionately much larger than a comparable battery in GEO delivering the same Wh energy during each discharge.

It is noteworthy from Fig. 10.12 that the life at given temperature is an inverse function of the depth of discharge. If the life is 100 units at 50 % DoD, then it would be about 200 units at 25 % DoD. The *cycle life* × *DoD* product is roughly constant in the first approximation, although it decreases with increasing temperature. Such is true for most electrochemistries. This means that the battery at a given temperature can deliver the same number of equivalent full charges regardless of the depth of discharge. Phrased differently, the total Wh energy that the battery can deliver over its life is roughly constant. The battery lasts proportionately longer if less energy is used per cycle. This observation is useful in comparing the mass and cost of various battery options for a given application at the conceptual design stage.

Once the electrochemistry and the number of parallel batteries are settled, the battery design depends on system parameters such as (1) bus voltage and load current, (2) charge and discharge rates and duration, (3) operating temperature during charge and discharge, and (4) life in terms of number of charge and discharge cycles. The life consideration is the dominant design driver in setting the Ah ratings. Even when the load may be met with a smaller capacity, the battery is oversized to meet the cycle life requirement. For example, with the same Wh load, the battery that must deliver twice as many cycles approximately double the capacity.

The issue of in-orbit battery reconditioning is considered in Chap. 20.

10.2.3 Power Electronics

Major power electronic components used in the spacecraft are (1) shunt regulator for bus voltage control during sunlight, (2) battery charge converter (buck converter), and (3) battery discharge converter (boost converter). They control the bus voltage and convert the voltage to match the operating voltages of various components. The voltage conversion is performed by solid-state semiconductor devices used as controlled switches which are turned on and off at high frequency. Capacitors and inductors are used to store energy when the switch is connected to the power source. The stored energy is then discharged to continue powering the load when the switch is off. Transformers are used where needed.

10.2.3.1 Switching Devices

A variety of solid-state devices are used as controlled switches. However, the devices commonly used in space are (1) metal-oxide semiconducting field effect transistor (MOSFET), (2) bipolar junction transistor (BJT), and (3) insulated gate bipolar transistor (IGBT). The device selection depends on the required voltage, current, and switching frequency. A common feature among these devices is that all are three-terminal devices. Their generally used circuit symbols are shown in Fig. 10.13. The two power terminals 1 and 0 are connected in the main power circuit. The control gate terminal G is connected to the auxiliary control circuit. In normal conducting operation, terminal 1 is generally at higher voltage than terminal 0. Since the device is primarily used for switching power on and off as required, it is functionally represented by a gate-controlled switch. In the absence of the gate control signal, the device resistance between the power terminals is large—the functional equivalence of an open switch. When the control signal is applied at the gate, the device resistance approaches zero, making the device function like a closed

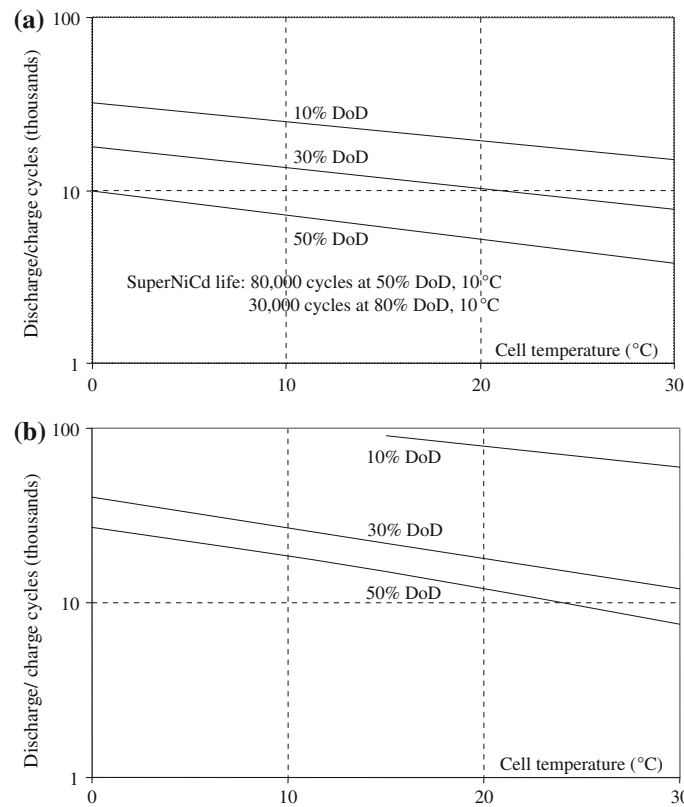


Fig. 10.12 Cycle life versus temperature and DoD for NiCd and NiH₂ batteries. **a** NiCd battery life. **b** NiH₂ battery life

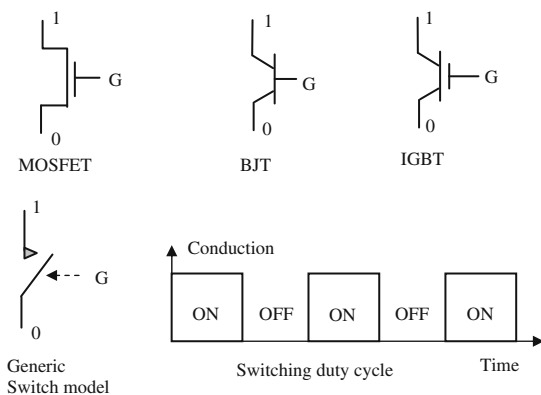


Fig. 10.13 Power electronics switching devices used in space

switch. The current through the switching device has a maximum saturation level regardless of the voltage applied between the power terminals 1 and 0.

The switch is triggered periodically on and off by a train of gate signals of suitable frequency. Within a separate triggering (firing) circuit, a sharp rectangular signal is derived by comparing a voltage control signal with a triangular or sawtooth waveform. This on–off drive is then applied to the gate of the semiconductor switch. Although the control circuit has a distinct identity and very much

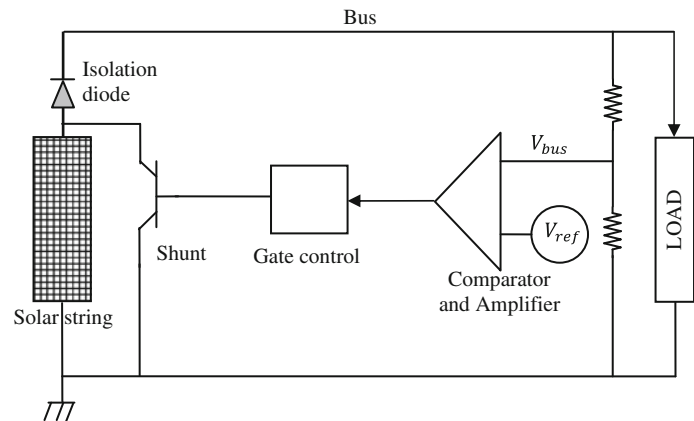
Table 10.4 Maximum voltage and current ratings of power electronics switching devices

Device	Voltage rating (V)	Current rating (A)	Remark
MOSFET	1,000	100	Offers higher switching speed, simpler firing circuit
BJT	1,500	200	Requires larger current signal to turn on
IGBT	1,200	100	Combines the advantages of BJT and MOSFET

different design features, it is often incorporated into the main power electronic component assembly. The transistor switch is turned on and off at high frequency, typically at 50–200 kHz, and sometimes higher. The duty ratio D of the switch is defined as $D = (time\ on/switching\ period) = (T_{on}/T) = T_{on} \times Switching\ frequency$.

The available voltage and current ratings of the switching devices and their gate triggering requirements vary with the device type. The presently available ratings are listed in Table 10.4, not all of which are space qualified. The power electronic components that use such high frequency switching devices are discussed next.

Fig. 10.14 Full shunt regulator for bus voltage control



10.2.3.2 Shunt Regulator

At the beginning of spacecraft life, the power output of the solar array during sunlight normally exceeds the load plus the battery charge requirements. The excess power must be diverted (shunted) from the bus in order to control the bus voltage. The shunt load can be a dump resistor, which would convert the solar array power into heat. Such heat dissipation in the spacecraft body would pose a burden on the thermal system in providing adequate cooling. An alternative commonly used in spacecraft is to shunt some of the solar array strings to the ground. This forces the string to operate under short circuit condition, delivering I_{sc} at zero voltage. In the shunt mode, no power is delivered to the bus or to the ground. The photon energy remains on the array, raising the array temperature and ultimately dissipating the excess power to space. The solar array is essentially being used as the thermal dissipater.

Figure 10.14 depicts a typical shunt regulator where a transistor is used as the switch. When the excess power is available, the bus voltage will rise above the rated value. This is taken as a signal to turn on the shunt switch across the required number of solar array strings. Thus, the shunt is turned on or off by a transistor controlled by the bus voltage reference. For an array with many strings in parallel, the basic configuration shown in Fig. 10.14 is used for each string separately. The same gate signal is supplied to all modules simultaneously in small power applications. For shunting large power, multiple shunt circuits are switched on and off in sequence to minimize the switching transients and the resulting electromagnetic interference to the neighboring equipment. For fine voltage control, the last shunt to turn on is operated in the pulse width modulation (PWM) mode, while all others are fully on or off.

Another application of the shunt regulator is in small satellites, where a dedicated solar array module is used to directly charge the battery without a battery charge regulator. When the battery is fully charged, the solar array module is shunted to ground by shorting the switch. This way, the battery is protected from overcharging.

10.2.4 Distribution Harness

The power distribution harness includes the insulated conductors, connectors, and the shield. Its mass is determined from the detailed layout and routing of all of the wiring required after the spacecraft has been well defined. For this reason, the harness mass is often considerably heavier than that estimated at the preliminary design stage. A typical harness mass breakdown is (1) 30 % in wires between power boxes, (2) 20 % in solar array wires, (3) 30 % in command and telemetry wires, and (4) 20 % in all connectors.

The wire size is measured in American Wire Gage (AWG) or in mm^2 cross section in metric wire gage. The AWG and Birmingham (BWG) numbers are inverse measures of the conductor's bare diameter, and are set on a log scale, i.e. $\text{AWG} = 20 \text{Log}(0.325/\text{diameter in inches})$. Thus, for every one gage up, the diameter increases by a factor of 1.1225 and the area by 1.26. The diameter doubles every six gages and the area doubles every three gages. The maximum current carrying capacity (ampacity) of the wire in space is less than that on the ground due to the absence of convective cooling. This requires de-rating the wire ampacity for space applications from the ground-based rating. The ampacity of various wires gages are listed in Table 10.5.

The most commonly used electrical conductor is copper for its good performance and low cost. Annealed copper has high conductivity but low tensile strength. For this reason, wires thinner than AWG 20 are often required to use high strength copper alloy 135, which has 40 % higher tensile strength, and 10 % higher electrical resistance. Copper wire coated with nickel or silver is used to resist corrosion and oxidation. Tin plated wires are widely used on the ground, but are forbidden in space due to the growth of whiskers.

Aluminum is sometimes used in power equipment where lightweight and/or low cost is desired. It is used in overhead transmission lines and pole mounted power transformers on the ground, and in some aircraft and commercial spacecraft

Table 10.5 Maximum allowable amperes in wires and connector pins of same gage

AWG ^a	Diameter (in.)	Single wire in free air on ground (MIL-STD-5088)	Wires in space ^b in 70 °C ambience (MIL-STD-975 and GSFC-PPL-19)	
			Single wire	Bundle or cable ^c
30	0.0100	n/a	1.3	0.7
26	0.0159	10.5	2.5	1.4
24	0.0201	14	3.3	2.0
20	0.0320	24	6.5	3.7
16	0.0508	37	13.0	6.5
12	0.0808	68	25.0	11.5
8	0.1285	135	44.0	23.0
4	7×0.0772	260	81.0	40.0
0	19×0.0745	460	147.0	75.0

^a Wires sizes AWG 10, 14 and 18 are not used in aerospace for general wiring, and AWG 2 and 6 have no counterpart electrical connector contacts (pins)

^b For TFE Teflon insulated wires rated for 200 °C

For 150 °C rated insulation, use 80 % of values shown

For 135 °C rated insulation, use 70 % of values shown

For 105 °C rated insulation, use 50 % of values shown

^c For cable bundles of 15 or more wires in 70 °C ambience in hard vacuum. For smaller bundles, the allowable current may be proportionately increased as the bundle approaches a single conductor

Table 10.6 Copper and aluminum conductor comparison

Characteristic	Copper	Aluminum
Resistivity, Ωm at 20 °C	1.724×10^{-8}	2.830×10^{-8}
Mass density, g/cm^3	8.89	2.70
Temperature coefficient of resistance α per °C	3.93×10^{-3}	3.90×10^{-3}
Melting point, °C	1,083	660
Flex life (relative)	1	0.5
Thermal coefficient of expansion (relative)	1	1.4
Creep rate at 65 °C (relative)	1	1,000

harnesses. The performance of aluminum is compared with copper in Table 10.6. For the same power loss or voltage drop, copper can be replaced with aluminum of relative mass equal to resistivity by mass density product ratios. Aluminum conductor, therefore, would theoretically have $(2.830/1.724) \times (2.70/8.89) = 0.50$ or 50 % of the copper mass for the same electrical performance. However, in practice, aluminum does not produce 50 % mass saving due to various mechanical reasons.

Insulation is designed to withstand the rated and abnormal transient voltages. The transient voltage can be higher by several times the rated value. The insulation design must preclude corona and arcing at pressures below 10 torr, and withstand the radiation environment and atomic oxygen. In high radiation spacecraft, such as GPS, the system specifications often require that silicon-insulated wire not be used and that the solar array wires and interconnects be welded to withstand high radiation.

A cable shield is wrapped around the wire bundle to prevent electromagnetic interference from entering the cable or radiating out. The shield can weigh 15–40 % of the cable weight. The shield options are braid versus tape, and copper versus aluminum. The braid is used when extreme flexibility is required. Its mass as compared to the wire conductor is about 40 % for flat cables and 20 % for round cables. Thin 2-mil (50 μm) copper vapor deposited on Mylar or Kapton tapes are widely used. The tape is applied on the cable with insulation touching the cable, followed by another layer on the top. The shield mass with such tapes is roughly 80 % of that with braid.

10.2.5 Solar Array Drive

The solar array drive and array drive electronics (SAD/ADE) together provide the capability of rotating the solar panels with respect to the spacecraft body. The operation of the SAD is controlled by decoded uplink commands from the on-board computer (OBC). It receives a timing clock and a synchronizing signal from the OBC. In the 3-axis stabilized 2-wing geostationary satellite, one axis is always aligned with the local normal to the Sun and another axis along the orbit normal. Two sets of open loop (clock controlled) solar array drive motors maintain the Sun orientation. A brushless DC stepper motor rotates each panel separately. The slip rings on the SAD shaft provide the interface between the rotating panel and the fixed Earth-pointing spacecraft body. One SAD controls the north panel and the other controls the south panel. The two are

interchangeable in design, where the ‘forward’ direction of rotation is separately selected by external means. Each SAD has only one mechanical assembly, but has redundant motor windings and redundant position telemetry potentiometers. Only one set of windings is powered at a time. The redundant windings are fully isolated to prevent failure propagation. The reliability of 0.99 is typical for both sides combined over a 15 year mission.

In an Earth orbiting satellite, the solar array is rotated once per orbit by the SAD to track the Sun at or near normal angle. The rotation is rate-servo controlled. The body information and position errors are computed by OBC to derive rate control signals. The nominal rate of rotation is mission specific and is primarily determined by the respective orbit characteristics. Using slip rings and carbon brushes is one way of providing the rotary joint between the rotating array and the satellite body. The control signals for the required rotation rate come from the telemetry, tracking and command (TT&C) system, which also selects the rotation direction. The mass of each SAD/ADE assembly can be 5–10 kg in GEO communications satellites.

The SAD/ADE provides telemetry defining its status and that of the solar panel. Each unit provides a potentiometer voltage signal which is directly proportional to the angular position of the panel shaft, ranging from 0 V at 0° to +5 V at 360°. The typical SAD uses a four-phase, 16-pole permanent magnet switched-reluctance stepper motor to drive a zero-backlash harmonic drive. Each phase coil resistance is in the 50–100 Ω range. A vanadium-cobalt steel stator core and a neodymium-iron-boron permanent magnet rotor provide high torque per unit mass. The rotor is typically on 440C stainless steel ball bearings and a titanium case. Each SAD/ADE draws power from the essential battery bus on a switchable and protected output line. The peak input power is about 10 W and the average power about 1 W. Switch-mode power pulses applied to the stator coils at a suitable frequency drive the motor. The grounding scheme for the ADE uses four separate grounds, one each for the power, signal and communications circuits, and one for all equipment chassis.

10.2.6 Electro-explosive Deployment

Electro-explosive deployment (EED) is the traditional deployment device. It is also known as pyro-technic ordnance. It uses electrical energy to ignite the squib of explosive powder. The resulting force deploys the solar array or other component, typically under a spring-loaded force. The EED will be ignited when an applied current imposes a certain amount of energy upon the squib within a specified time. For safety reasons, the squib must withstand

certain minimum energy without igniting. The typical EED is rated at 1A-1 W for no fire, and 4A-4 W for sure fire.

The EED requires heavy shielding and great care with regard to the electromagnetic interference (EMI) pickup. Moreover, the EED explosive is thermal and shock sensitive. Therefore, its installation is sometimes done at the launch pad, which costs much more than at the factory. Most manufacturers install EED squibs at the factory and transport the spacecraft as class 1 explosive (sensitive to thermal and shock environment), which is also expensive.

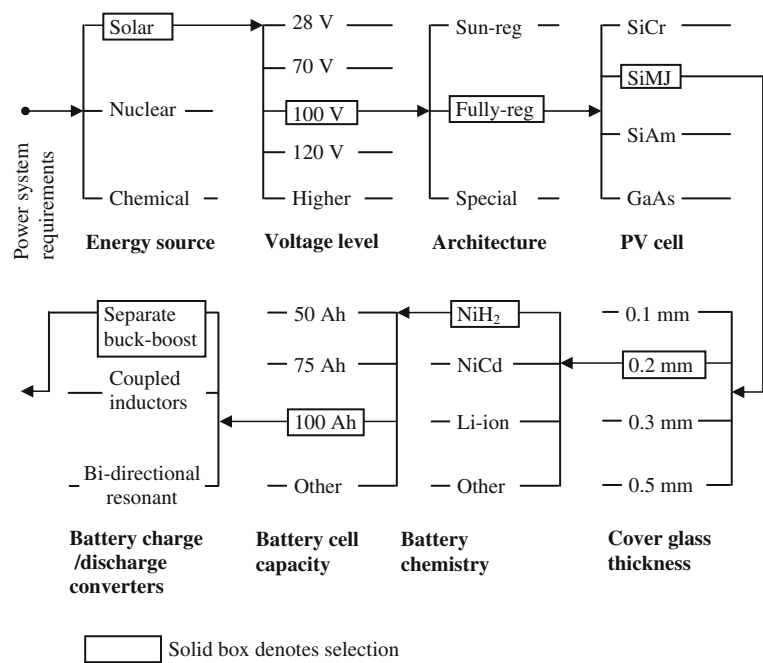
The EED harness is routed separately from the power and signal harness in order to minimize the EMI concerns for safety. The EED technology, although used for several decades, has some disadvantages, such as (1) all spacecraft components must be designed to withstand severe pyro shock excitation, and (2) high safety related costs in documenting and reporting alerts, and complying with all other stringent regulations. The newer EMI-free alternatives are laser initiated and use shape-memory metal for deployment.

10.2.7 Design Process and Trades

Once the spacecraft-level trades are settled, the power system engineer focuses on the internal EPS-level trades that may reduce the mass and cost. The power system mass as a percentage of the satellite dry mass can range from 25 % in LEO satellites to 45 % in GEO satellites. Saving even a few percent of power system mass can result in appreciable savings at the spacecraft level. The first task of the power-system design engineer is to select the optimum primary energy source.

The bus voltage level is selected based on the power level. Early spacecraft with loads of a few hundred watts used 28 V dc, which was primarily based on the product specifications readily available for the aircraft power system at the time. Since then, the power levels have increased significantly. With power being the product of voltage and current, a high power requires a high-voltage bus in order to keep the current level at a reasonable level. Otherwise, the excessive power loss in switching devices and I^2R loss in conductors reduce the system efficiency considerably. Today’s spacecraft bus voltages, somewhat standardized by the product lines of various manufactures are 28, 50, 70, and 100 V. The ISS has 160 V DC solar array voltage and 120 V DC distribution voltage [6]. The 160 V limit comes primarily from the bare conductor interaction with space plasma, particularly in LEO. Above 160 V, the solar array current leakage to plasma increases exponentially with potential sparking above 180–200 V.

Fig. 10.15 Power system design trades flow-chart



Voltages higher than 160 V can be used in low Earth orbit with insulated cables covered in a shielded enclosure, and by encapsulating (grouting with insulating compound) all solar cell edges, connectors, and circuit board. For early space station designs [6], NASA considered 120 V dc, 270 V DC and 440 V 20-kHz ac. It finally selected 160 V DC for solar array voltage and 120 V DC for distribution bus, with necessary step-down converters for existing 28 V DC hardware. For any spacecraft, the influence factors in the voltage selections are (1) power level as the primary driver, (2) space environment and space plasma, (3) the Paschen minimum breakdown voltage between bare conductors, (4) human safety, and (5) availability of components, such as semiconductor devices, power distribution and protection devices, tantalum capacitors, etc.

Next, the power generation and energy storage technologies are jointly selected to optimize the total power system. The major driving factors are the payload power level, the operating orbit, mission life, number of satellites in the program procurement, etc. For the self-derived satellite-level load requirement, the trade study is done to select various key component for the power system, such as the PV cell, cover glass thickness for radiation protection, array substrate, battery electrochemistry and cell Ah rating, power converter topologies, etc. An example of such trades is displayed in Fig. 10.15. As the trade study proceeds from left to right, the selections made are shown by continuing arrows. Since the solar array and the battery are two components that primarily contribute to the power system mass and cost, they get more attention and see more rapid technological changes than other components. They impact not only the EPS design, but also other spacecraft systems.

10.2.8 Power System Requirements

The electrical power system requirements are seldom found in the customer specifications for the spacecraft under procurement. They are derived from the spacecraft-level requirements and in-house trade studies. The EPS self-derived requirements are based on various analyses, but the final requirements generally come from the operational orbit analyses. However, the transfer and other orbits must also be analyzed to ascertain that the proposed requirements are met in the worst case. Major self-derived requirements of the power system are (1) solar array EOL power level, (2) solar array pointing and rotation for Sun orientation, (3) battery Ah capacity, (4) battery DoD and charge control, (5) bus voltage regulation, and (6) EMI, EMC and electrostatic discharge, ESD.

The power-system design team performs the following worst-case analyses to establish detailed EPS requirements

- Power flow to determine the component ratings and heat dissipations
- Energy balance to determine the battery rating
- Voltages at the terminals of various equipment
- EOL and BOL solar array power generation capability
- Load switching and fault response, including major fuse-clearing events
- Bus stability under various feedback control loops
- Energy balance in operational and transfer orbits
- dV/dt specifications, which come from three sources
 - voltage fall after a short circuit fault until the fuse clears
 - voltage rise after the fuse clears
 - sudden loss of power during integration and testing.

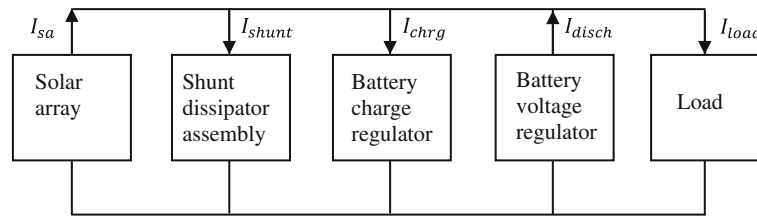


Fig. 10.16 Power system components participating in energy balance

10.3 Power System Performance

10.3.1 Energy Balance and Power Management

By designing the spacecraft electrical power system, what is really meant is designing the electrical energy system. The satellite has a limited time in orbit to generate power, but the loads need to be powered all the time. The battery stores energy during sunlight and delivers it to the loads during eclipse. The energy balance between the battery charge and discharge over one orbit period must be on average positive with some margin. Otherwise, the battery would walk to total depletion in a matter of time. The power to and from various components must therefore be managed in order to maintain the energy balance in both the transfer orbit and the operational orbit.

The energy balance analysis is performed at the design stage by simulating the power flow and energy account on a computer. The program is generally structured to allow analyses on the baseline design and its derivatives, and to answer many *what if* questions in normal and abnormal operation. Such analysis is performed during all phases of the mission for a given launch date (year, month and day). Various fault conditions, including battery cell failures (short or open) and loss of a solar array circuit, are simulated to determine energy balance under the worst-case condition(s). Other equally important uses of the energy balance computer program are

- Determine and/or optimize the load capability of a given EPS
- Derive component ratings based on maximum power flow in each component
- Determine power dissipation for thermal design of each component, particularly the battery, since its performance is highly temperature sensitive.

The computer program for such analysis is generally developed around variable parameters with no *hard-coded* numbers. This allows greater flexibility in using the tool for a wide variety of applications. Figure 10.16 depicts EPS components contributing to the energy balance program. The basic equations for currents and battery DoD that are

computed typically every second or so in the energy balance analysis are

$$\text{During sunlight, } I_{sa} = I_{load} + I_{chrg} + I_{shunt} - I_{disch} \quad (10.6)$$

$$\text{During eclipse, } I_{load} = I_{disch} \quad (10.7)$$

$$\begin{aligned} DoD(t) = & DoD_0 \\ & + (\text{Sum of Ah delivered} / \text{Actual Ah capacity}) \end{aligned} \quad (10.8)$$

where DoD_0 is the initial DoD.

The entire program is divided into several software modules, each representing various components. Due to the non-linear nature of the battery cell and solar array performance parameters, the program typically uses *static* lookup tables to determine the cell performance characteristics as a multi-variable function of the battery current, temperature, state of charge, and the solar array operating voltage. Programming with a computer language, instead of modeling on a spreadsheet, significantly improves the capability of the program.

10.3.1.1 Dynamic Performance and Stability

The dynamic bus impedance and the control loop gain influence the dynamic performance of the power system under an internal or external transient perturbation. Key performance attributes coming out of the dynamic study are the bus voltage ripples, transient deviations, fault and fuse-clearing transients, and the control loop stability under harmonic ripple excitation. On the other hand, the static performance under a slow change or after the dynamic response has settled, is largely influenced by the static impedance of the bus. The bus voltage change long after a load change is an example of static performance. Since the dynamic and static bus impedances are similarly defined, they are jointly covered in the following section.

10.3.1.2 Bus Impedance

A complex electrical network having a number of sources and loads between any two terminals can always be reduced to a simple Thevenin equivalent source model consisting of one source voltage V_s with an internal series impedance

Fig. 10.17 Thevenin equivalent source model of complex electrical network. **a** Complex electrical network. **b** Thevenin equivalent source

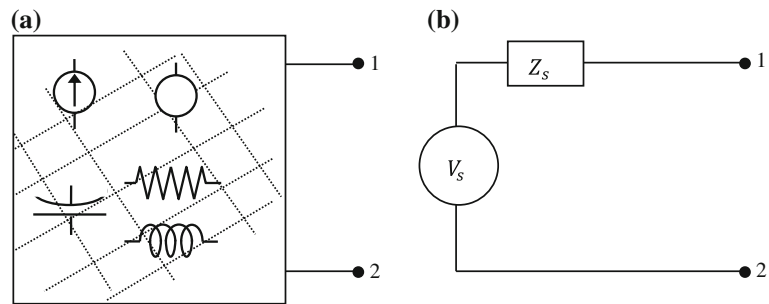
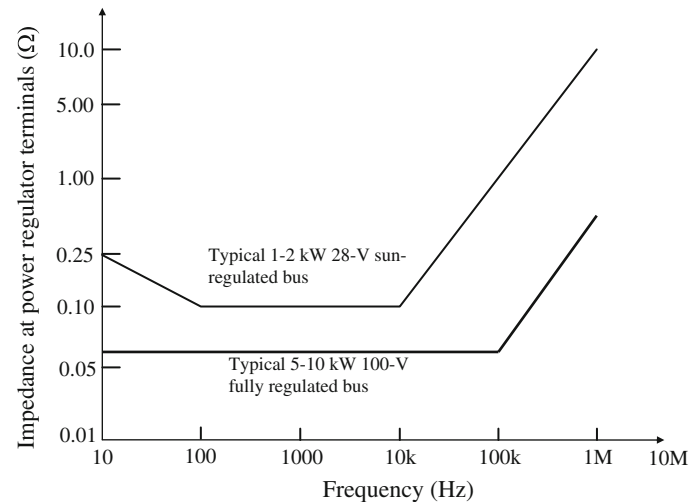


Fig. 10.18 Main bus source impedance for a 28 and 100 V GEO communications satellite



Z_s (Fig. 10.17). The two source parameters are determined as follows

- With open circuit between load points 1 and 2 as in Fig. 10.17b, but with all other parameters at rated values, the voltage between terminals 1 and 2 equals the source voltage V_s (since the internal voltage drop is zero under zero load current). Therefore, V_s is the open circuit voltage of the system at the load terminals.
- With the terminals 1 and 2 shorted, the internal voltage is now totally consumed in driving the current through the source impedance only. Therefore, Z_s is the open circuit voltage/short circuit current at the load terminals.

The short circuit current can be determined by calculations, or by tests at a reduced voltage applied to limit the current to the rated value. The full short circuit current is then calculated by scaling to full rated voltage. Any non-linearity, if present, must be accounted for. The source impedance of most spacecraft bus architectures is highly non-linear due to use of multiple bus regulators along with dead-band regions. The dynamic performance of such systems is largely driven by the transient nature of switching from one mode to another, such as solar array shunt control to battery discharge control. Linear modeling may be acceptable for steady-state regulation and small signal load changes within the control range of each controller, but not

for modeling mode change transition, which must use a transient model for the individual control.

The Thevenin equivalent source model derived under the steady-state static condition gives the static bus impedance Z_s . The source impedance derived under the dynamic condition (that is for an alternating or incremental load) is the dynamic bus impedance Z_d . This varies with frequency and can be either calculated or measured by test. With the bus in operational mode delivering its rated load, a small high frequency AC current I_h is injected into the bus using an independent current source, and the value of V_h , the high frequency voltage perturbation in the bus voltage, is measured. The dynamic bus impedance at that frequency is then $Z_d = (V_h/I_h)$. Since Z_d has a strong influence on the system's dynamic performance, it is kept below specified limits. Figure 10.18 is a typical main bus source impedance for a mid-size GEO communications satellite.

10.3.1.3 Stability Criteria

The steady-state stable operating point is where a PV panel's power output equals the load power. The constant power load has two such points, A_1 and A_2 shown in Fig. 10.6. If point A_1 gets a small disturbance of $+\Delta V$ for any reason, there would be a positive power excess, moving the voltage higher from A_1 , and again further higher in a runaway

situation. On the other hand, at point A_2 , a small voltage disturbance of $+\Delta V$ due to any reason would result in a power deficit, restoring the system back to its original operating point A_2 . Thus, point A_2 is a stable operating point, where the system works like a negative feedback system. Analytically, an operating point is stable if and only if

$$\left[\frac{dV}{dI} \right]_{source} < \left[\frac{dV}{dI} \right]_{load} \quad (10.9)$$

In terms of the absolute values of the dynamic impedances, Z_s is the source output impedance and Z_L is the load input impedance, the system is stable only if $Z_s < Z_L$ at all frequencies. When $Z_s > Z_L$, the system is not necessarily unstable and requires further analysis to determine the stability. The constant power load gives a stable operating point only on the right-hand side of the P_{max} point. The solar array powering a purely resistive load is always stable, since Eq. 10.9 is always true.

10.3.2 Electromagnetic Interference, Compatibility and Electrostatic Discharge

All spacecraft systems are required to be compatible with the interference expected from internal and external sources. For decades, the electromagnetic interference (EMI) and electromagnetic compatibility (EMC) requirements have come from MIL-STD-461. It specified the maximum emission limit of the potential culprit, and the minimum susceptibility level of the potential victim equipment. The companion MIL-STD-462 defined the test methods for verifying that MIL-STD-461 requirements were met, and MIL-STD-463 defined the applicable terms and units. The first two standards are now merged into one, MIL-STD-461. The contractor of commercial and defense spacecraft is required to develop three documents and submit to the customer as deliverables: (1) EMC control plan, (2) EMI test plan, and (3) EMI test report.

10.3.2.1 EMI Sources and Suppression

The EMI requirements broadly fall in two general groups, the conducted EMI and the radiated EMI. In fact, EMI can enter the equipment either by conduction via wires, or by radiation in space. In verifying that a spacecraft will meet these requirements, the first order of task is to determine the conducted EMI and the radiated EMI from potential sources, and the degree of coupling to the victim equipment. In space systems, the main sources of EMI are (1) switching large current or voltage at high frequency causing large (dI/dt) and (dV/dt) , (2) electrostatic discharge, and (3) nuclear detonation around the spacecraft. Various

methods of controlling and/or suppressing the EMI in the spacecraft are, in the order of their importance

- Minimize the EMI generation in the first place by
- Minimizing the current loop area in switching circuits
- Minimizing the switching transient's (dI/dt) rate in large current loops
- Using snubber capacitors to minimize the voltage transient's (dV/dt) rate
- Minimize the E and B field couplings between the culprit and the victim equipment by
 - Minimizing the inductive coupling by twisting wires or using coax cables
 - Minimizing the capacitive coupling by using shields and by reducing area of exposed metal and keeping it far from the ground, since $C = (kA/d)$.
- Divert the energy impinging on the victim equipment to ground by using
 - Proper grounding scheme
 - Faraday shield, single or double.
- Protect the equipment from the coupled energy by using
 - L-C filters for conducted EMI. Enclosure shield for radiated EMI.

10.3.2.2 Electrostatic Discharge

An electrostatic charge accumulates on any probe in space regardless of its being in or out of the van Allen belts. The accumulated charge raises the electrical potential of the probe, causing a current flow from the probe to the surrounding plasma. If the current cannot maintain a balance of charge, the probe's potential will keep rising until arcing takes place. This problem can occur particularly when the spacecraft leaves or enters an eclipse, when the interaction with space suddenly changes. Arcing can also arise due to differential charging of insulated surfaces that are not electrically connected. Each isolated surface acts as an independent probe in space, which eventually floats to a potential that results in no net current to or from the space plasma. That potential is of the order of the plasma kinetic energy. Insulating surface do not distribute surface charge, hence can charge up to much higher differential potential until discharge takes place by way of arcing and/or flashover.

The charging, the subsequent electrostatic discharge (ESD), and their remediation in GEO and in LEO are significantly different. In the high-energy plasma environment of GEO, the electron charge that accumulates on insulating surfaces increases the electric field to adjacent conductors above the breakdown level, leading to arcing. The resultant arcing currents traveling through conductors can upset electronic components and induce spurious signals. A common design solution for GEO is to coat all outside surfaces of the spacecraft with conducting materials. This prevents differential charging by distributing the charge over all

surfaces and equalizing their potentials. A coating material having surface sheet resistivity of less than $5 \text{ k}\Omega/\text{sq}$ ¹ is considered adequate to eliminate differential charging. On solar arrays, indium oxide type conductive and transparent coating is applied on the cover glass when required.

In LEO, because of the high thermal plasma current density, surfaces do not ordinarily collect much differential charge. The major concern in LEO is the absolute charging of spacecraft surfaces with respect to the surrounding plasma. Normally, the collected plasma current bleeds off the absolute potential rapidly.

10.3.3 De-rating Parts for Reliability

The U.S. Military Handbook-217 (MIL-HDBK-217) establishes the uniform method of predicting the reliability of military electrical and electronic parts, equipment, and systems. It lists the base failure rates per million hours of operation for numerous parts under base electrical, thermal, and mechanical stresses. Any deviations from the specified operating conditions would alter the failure rate listed in the handbook.

The failure rate of an electrical component depends primarily on voltage and temperature. The electrical insulation at high temperature oxidizes, becomes brittle, and may crack, leading to failure (short circuit). The oxidation is a chemical degradation, and follows Arrhenius exponential growth. Data from a range of electrical equipment shows that the failure rate doubles (or the life is shortened to one-half) for every $7\text{--}10^\circ\text{C}$ rise in the operating temperature. In the reverse, the life doubles for every $7\text{--}10^\circ\text{C}$ reduction in the operating temperature. Similar degradation (wear) takes place above a certain voltage, although it is not as well understood as that for the temperature. The rise in failure rate at high stress level is not to be confused with the wear-out failure rate. It raises the flat part of the classical bathtub curve for reliability. The failure rate is still constant per unit time, although another constant at another operating stress level.

De-rating is the reduction of electrical, thermal, and mechanical stress levels applied to a part in order to decrease the degradation rate and prolong the expected life. It is routine practice to decrease the wear-out failure rates in

military and space worthy designs. The de-rating in current is often done to lower the temperature. On the other hand, the current de-rating in some active devices, such as transistors, is done to control the (dI/dt) rate, which can upset the semiconductor operation; de-rating increases the margin of safety between the operating stress level and the actual failure level of the part. It provides added protection from system anomalies unforeseen by the design engineer.

Most space programs maintain their own preferred parts list based on the failure rates. This preferred parts list also gives de-rating factors, which must be applied to all designs on those programs. Both the selections of parts and the de-rating factors in the preferred parts list are based on the heritage designs successfully flown. Space agencies, such as NASA and ESA, maintain their own preferred parts list. If a desired component is not in the preferred parts list, it must be qualified by rigorous testing under the same environment as those in the list.

10.4 Special Power Systems

10.4.1 Interplanetary Mission

The mission environment depends heavily on the spacecraft's distance from the Sun. For interplanetary missions away from the Earth—either, closer to the Sun or farther away from the Sun—the solar array, battery, and power electronics designs differ significantly because of the significantly different environment encountered. The extreme temperature on either side—high or low—has a large impact on the performance. The solar flux at any distance in deep space is given by $I = (I_{earth}/R^2)$, where I_{earth} is the solar flux in the Earth's orbit, and R is the distance from the Sun in astronomical units (au). This equation assumes the Sun to be a point source, and may give some error at distances less than a few solar radii. The solar array power output varies linearly with the incident solar flux. Therefore, as the spacecraft moves away from the Sun, the power decreases inversely with the distance squared. The PV array temperature also decreases in the same ratio, which results in a higher PV conversion efficiency. The combined effect of the flux and the temperature changes is such that the photovoltaic power generation varies not inversely with the distance squared, but with R^α , where α is approximately 1.5. Table 10.7 lists planets in our solar system with their distances from the Sun in au, and approximate power generation rate in their orbits, considering both the solar flux and the temperature variations. See also Table 4.3.

Due to the Sun's proximity, the electrical power system for a Mercury mission must meet the harsh thermal and radiation environment. Haines [7] of the European Space Agency has reported the power system design for a Mercury

¹ Sheet Resistance is a special case of resistivity for a sheet of uniform thickness. The SI unit of resistivity is ohm · meter ($\Omega \cdot \text{m}$), which is more completely stated in units of $\Omega \cdot \text{m}^2/\text{m}$ ($\Omega \cdot \text{Area}/\text{Length}$). When divided by the sheet thickness, $1/\text{m}$, the units are $\Omega \cdot (\text{m}^2/\text{m}^2) = \Omega$. The alternate, common unit is 'ohms per square' (denoted ' Ω/sq '), is dimensionally equal to an ohm and exclusively used for sheet resistance avoiding misinterpreted as bulk resistance of 1Ω . Note that a square sheet with sheet resistance $50 \Omega/\text{sq}$ has an actual resistance of 50Ω independent of the size of the square.

Table 10.7 Solar flux and PV power generation in orbits of various planets in our solar system (relative to those in the Earth's orbit)

Planet	Distance from the Sun (au)	Solar flux relative to Earth orbits	PV power generation accounting for temperature difference
Mercury	0.31–0.47	10.40–4.52	Severe loss of voltage
Venus	0.72	1.93	1.63
Earth	1.0	1.0	1.0
Moon	1.0	1.0	1.0
Mars	1.66	0.36	0.59
Jupiter	5.20	0.037	0.084
Saturn	10.08	0.0098	0.031

sample return mission. It consists of three independent power systems for each phase of the mission. For example, the 20 kW, 100 V high-power system shown in Fig. 10.19 is used for electric propulsion, and is jettisoned just before the orbit insertion. After that, the 500 W, 28 V system is used for the orbiter, and a smaller power system for the surface landing, sample collection, and return phase of the mission.

10.4.2 Near-Sun Mission

The PV power system design for near-Sun missions between Mercury and the Sun needs special considerations due to the high temperature. The solar intensity increases to 100 Suns at 0.1 au (21 Sun radii, 1 Sun radius equals 0.00476 au), and to 2,500 Suns at 0.02 au (about 4 Sun radii). The PV cell loses power generation capability at such temperature due to loss in the open circuit voltage. Various options to limit the temperature to below 1,000 °C include reliably guaranteed array tilting, adding mirrors on the surface to decrease absorptivity and increase emissivity, partially silvered cover glass, and various louvers and shades to control the solar flux. Moreover, the PV cell having high band gap is needed. Figure 10.20 due to Brandhorst and Chen [8] shows effective power output as function of au distance and band gap of various PV cells operating below 1,000 °C. At distance greater than 0.5 au, the band gap has no significant effect on power generation. At distance less than 0.5 au, the higher band gap PV cell generates more power up to 0.1 au. Closer than 0.1 au, the PV cell becomes useless. The curves assume that the cell temperature is limited to 1,000 °C in all cases.

It is apparent from Fig. 10.20 that the PV power system has limitations in approaching the Sun at a close distance. An alternative approach may be to use the thermo-photovoltaic (TPV) direct energy conversion. There are several advantages, including easy coupling to a thermal source operating above 2,000 K. The feasibility of such an

approach has been demonstrated under US Department of Energy funding, but not fully developed.

Another alternative is to use a thermoelectric (TE) converter with the Sun as the heat source. Such a system is feasible for solar probes requiring instrument power under a few hundred watts. For example, NASA/JPL in 2003 designed the Sun-TE power system for a flyby probe to Jupiter and then towards the Sun to study coronal heating and the origin and acceleration of the solar wind. In the power system design reported by Choi [9], the probe's distance from the Sun varies greatly, from 5.2 au near Jupiter (gravity assist orbit) to less than 0.1 au (4 solar radii) near the Sun. The corresponding solar flux varies over 5 orders of magnitude from 50 W/m² to 4 × 10⁶ W/m². The spacecraft bus is shaded by the primary sunshield blocking the Sun. The shield's outside temperature is estimated to be 2,100 °C at 4 solar radii. A high temperature multi-layer thermal blanket keeps the spacecraft components cool. The shield and the blanket are made of carbon-carbon composite.

10.4.3 Deep Space Mission

Deep space and outer planetary missions cannot effectively use photovoltaic power generation due to insufficient solar flux. For those missions, an on-board radioactive isotope is often used to generate electrical power. The radioisotope heat is directed at a TE junction, which generates electrical potential just as in a thermocouple. The power system for such missions, therefore, typically includes a radioisotope thermoelectric generator (RTG), power electronics, and a small battery located inside the spacecraft body. The RTG heat may be sufficient to protect the power system from cold temperatures. If not, an additional isotope heat is needed to keep the electronics at required temperature. For example, unheated interplanetary spacecraft launched to explore the rings of Saturn would experience an average temperature of about −190 °C, which is the temperature of liquid nitrogen. For this reason, low temperature power electronic circuits have potential of finding applications in deep space missions. Such circuits designed and operated at low-temperature may result in more efficient system layout than the room temperature circuits. The advantages include reducing or eliminating the thermal shutters and the need for an isotope heat, which can cause overheating during launch. Understanding the performance of power electronics at extreme low temperatures is needed for this purpose. The following is known about the operation of power electronic components near the liquid nitrogen temperature.

Performance of certain semiconductor devices improves with decreasing temperature down to liquid nitrogen

Fig. 10.19 Main bus power system architecture for electric propulsion to Mercury. Image J.E. Haines, ESA

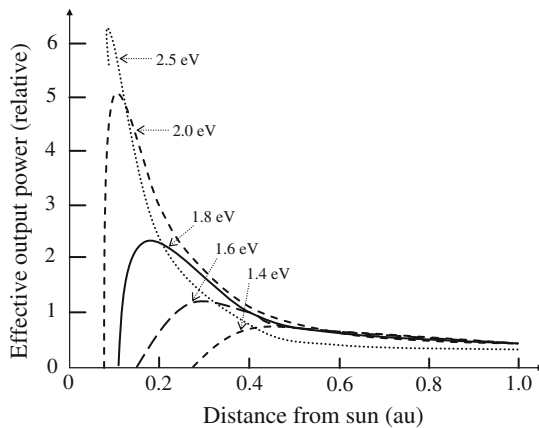
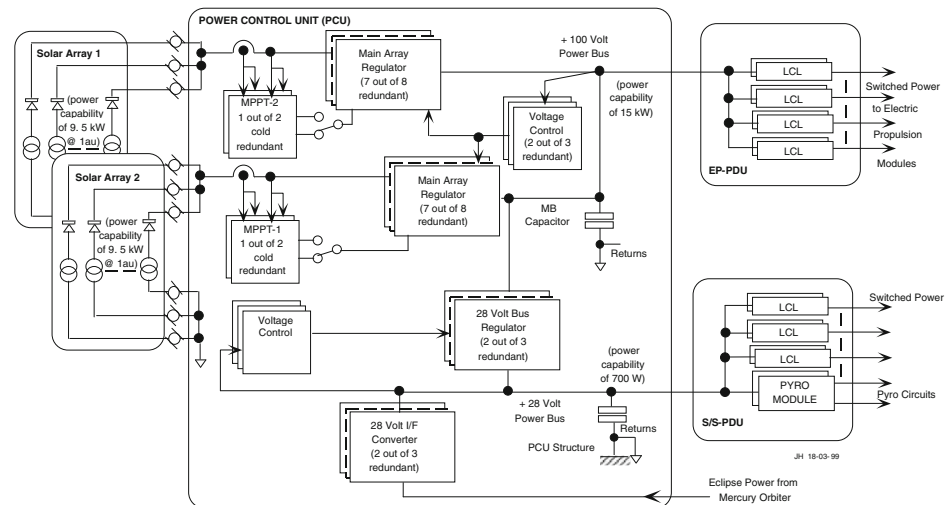


Fig. 10.20 Power output versus distance for PV cell of various band gaps at temperature below 1,000 °C

temperature. At low temperatures, majority carrier devices demonstrate reduced leakage current and reduced latch-up susceptibility. In addition, these devices show higher operating speed resulting from increased carrier mobility and saturation velocity. An example is the power MOSFET, which has lower conduction loss at low temperatures due to the reduction in drain-to-source resistance $R_{ds(on)}$ resulting from increased carrier mobility. NASA has tested other components such as resistors, capacitors, and magnetics that are needed for various power converters at liquid nitrogen operating temperature. Many of them have been found suitable for operating an unheated interplanetary spacecraft [10].

The battery can be a roadblock at very low temperatures. The Li-ion battery offers a somewhat favorable combination of energy and power density. However, its low temperature performance below -40 °C is poor. Tests have shown the following about the Li-ion cells [11, 12]

- Between room temperature and -20 °C, variations in electrolyte resistance and the anode to electrolyte

resistance are negligible, but the cathode electrolyte interface resistance increases substantially.

- The cell voltage and Ah capacity fall to approximately one-half at -40 °C. As a result, practically no energy was delivered at -40 °C. This is due to substantial increase in the total internal resistance.
- Poor cell performance at low temperatures can be attributed to the electrolyte becoming viscous or solid [13]. It is also attributed to the poor lithium diffusivity in the electrolyte. Work is underway to improve the low temperature performance of the Li-ion cell.

10.4.4 Radioisotope Thermoelectric Generator

An RTG for power levels of several hundred watts has been fully developed and used for decades. Such a power source has the advantage of supplying power all the times, thus eliminating the need for a battery in a base load system having no peak power requirement. An obvious disadvantage is the heavy radiation shielding required around the electronic components. The advantages of the RTG are

- It provides power for a long period, independent of the spacecraft orientation and distance from the Sun.
- It is suitable for missions far away from the Sun, too close to the Sun, or lunar missions with long eclipse periods.
- The power output is not affected by radiation damage in the Van Allen belts or from man-made nuclear threats.

The RTG consists of numerous thermoelectric cells connected in series-parallel combination to obtain the required voltage and current. Each TE cell converts the isotope thermal energy into electrical energy. The power conversion efficiency of the RTG depends on the material properties and the hot and cold junction temperatures T_{hot}

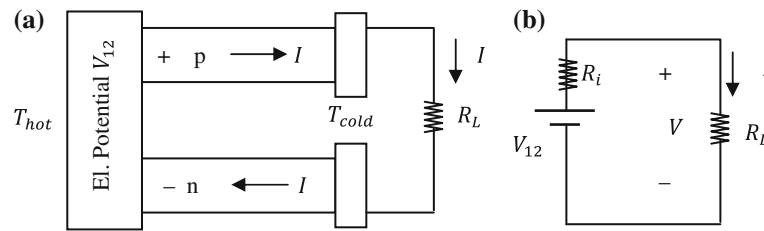
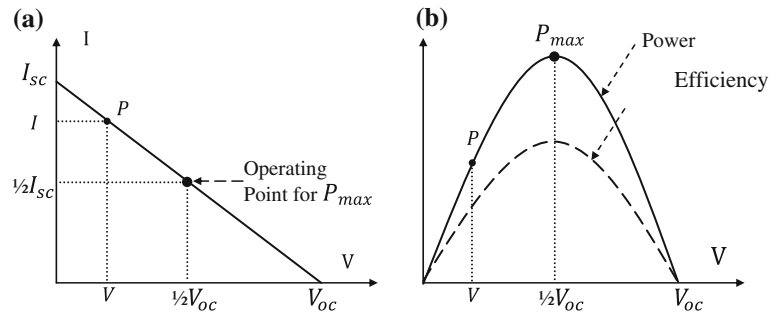


Fig. 10.21 Single-stage unicycle thermoelectric junction. **a** Construction **b** Electrical model

Fig. 10.22 Thermoelectric I–V and P–V curves with P_{max} point. **a** I–V characteristic. **b** P–V characteristic



and T_{cold} . The theoretical limit on this efficiency is Carnot cycle efficiency, $\eta_{carnot} = [(T_{hot} - T_{cold})/T_{hot}]$, where the temperatures are on the absolute Kelvin temperature scale. Practical designs yield about one-half the theoretical maximum efficiency. The most widely used material (Plutonium-238, ^{238}Pu isotope with SiGe TE cells) gives about 7 % conversion efficiency. Removing the remaining 93 % of the system energy as waste heat poses a significant design challenge. The specific electrical power output of RTG is typically low. Based on the total power system mass, it was 5 W/kg in the Galileo spacecraft. The ^{238}Pu isotope is also scarce and expensive, a few million dollars per kilogram. The US did not produce it for some time, instead purchasing it from Russia, with the US Department of Energy inventory of ^{238}Pu dropping below 10 kg, however US production of ^{238}Pu recommenced in 2013/14.

10.4.4.1 Thermoelectric Basics

The working principle of a TE converter is based on the Seebeck effect, which generates electrical potential when any two dissimilar materials are maintained at different temperatures. It involves electron or hole transfer between two dissimilar materials under thermal energy. The two materials can be conductors or semiconductors. The TE cells for space power applications use semiconducting materials, one p-type and the other n-type, as shown in Fig. 10.21. If two such dissimilar materials are held at a temperature difference $\Delta T = (T_{hot} - T_{cold})$, an electric potential difference V_{12} is produced at their junction. It is given by $V_{12} = (\alpha_{12}\Delta T)$, where α_{12} is known as the Seebeck

coefficient of the couple, generally expressed in $\mu\text{V}/^\circ\text{C}$. The coefficient α_{12} is often called the TE power, although it is not really a power. It is a characteristic constant, which depends on the material properties. The α_{12} is considered positive if the Seebeck voltage polarity produces current in the p-type material from high temperature to low temperature.

The total voltage generated due to the Seebeck effect works as an internal voltage source. With open circuit (zero load current), the external terminal voltage V is same as V_{12} generated internally. This voltage is designated as the open circuit voltage V_{oc} . When electrical current is drawn by load resistances R_L , there is an internal voltage drop. This is represented by an internal resistance R_i , which is approximately constant at a given temperature. The external terminal voltage V therefore decreases linearly with increasing load current, i.e. $V = (V_{oc} - IR_i)$. With the external terminals shorted, the maximum current flows to the load. This current is designated as I_{sc} , which is given by $I_{sc} = (V_{oc}/R_i)$. These equations can be rearranged to write $I = (I_{sc} - \gamma V)$, where $\gamma = (I_{sc}/V_{oc})$, the characteristic admittance of the RTG power source. The last equation in this paragraph gives the I–V characteristic of the RTG. It is a falling straight line from I_{sc} at zero voltage to zero current at V_{oc} as shown in Fig. 10.22a.

10.4.4.2 Maximum Power Extraction

The power transferred from the RTG to the load at any operating voltage V and load current I is $P = VI = V(I_{sc} - \gamma V) = (VI_{sc} - \gamma V^2)$. The power system design for

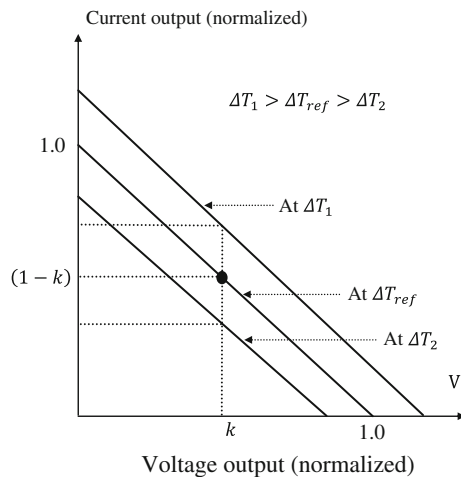


Fig. 10.23 Thermoelectric current versus voltage and various temperature gradients

extracting the maximum power from the RTG to the load must operate at a voltage such that $dP/dV = I_{sc} - 2\gamma V = 0$ at the operating voltage. This equation gives the corresponding operating voltage $V_m = I_{sc}/(2\gamma)$, which is also $1/2 V_{oc}$, and the current at that voltage is $1/2 I_{sc}$. The maximum possible power transfer from the RTG to the load is therefore

$$P_{max} = \frac{V_{oc} I_{sc}}{2} = \frac{V_{oc} I_{sc}}{4}. \quad (10.10)$$

P_{max} occurs when the RTG is operated at a voltage equal to one-half of the open circuit voltage, as shown in Fig. 10.22b. The power at zero voltage is obviously zero. It is also zero at V_{oc} since the current is zero there. In between, the power rises with the operating voltage, reaches the maximum value P_{max} , and then falls to 0 at V_{oc} . The RTG conversion efficiency is maximum at the maximum power transfer point.

10.4.4.3 Effect of Temperature and Aging

The I–V line of the RTG shifts upward for a higher ΔT , and downward for a lower ΔT , as shown in Fig. 10.23. The amount of shift is a characteristic of the couple material. Aging has a small effect on the RTG output, because the basic heat source has a long half-life in decades. For this reason, the power generation degrades a little, about 1.5 % per year (Fig. 10.24). Most power degradation is due to slow precipitation of the phosphorous doping in the n-type leg of the thermocouple. The I–V and P–V curves shift uniformly with time and temperature such that the maximum power point remains at the same voltage. This is a happy coincidence for the design engineer. The conversion efficiency is a function of the contact resistance and the hot and cold-side temperatures.

10.4.5 Dynamic System with Alternator

Solar energy can be used in a system other than photovoltaic. A dynamic energy conversion system is an example, where the Sun's energy is collected in the form of heat using a concentrator. The heat in turn is used to produce steam and drive a rotating turbo-generator or a reciprocating alternator to generate electrical power. Such a system was a primary candidate for the space station design in the 1980s for a power requirement of 300 kW. The system configuration is shown in Fig. 10.25. A parabolic concentrator focuses the Sun's heat on a receiver, which boils a fluid. The fluid can be a suitable liquid metal, such as potassium chloride. High-pressure steam of liquid metal produced in the receiver drives a turbine based on a Rankin cycle. The fluid can also be a gas, such as a mixture of helium or xenon having a molecular weight of around 40. The heated compressed gas in this case drives a turbine working on a Brayton cycle. A gas-based system minimizes erosion and sloshing problems in transporting the liquid metal. In either a liquid metal or a gas-based system, the high-pressure high-temperature fluid drives the turbine, which in turn drives an electrical generator. Waste heat transferred to the liquid coolant is dissipated via radiator panels to space. The energy conversion efficiency is much higher than the photovoltaic system. This minimizes the deployed collector area and the aerodynamic drag of LEO.

The usable energy extracted during the thermodynamic cycle depends on the working temperatures. The maximum thermodynamic conversion efficiency that can be theoretically achieved is Carnot cycle efficiency. Higher hot-side working temperature and lower cold-side exhaust temperature results in higher efficiency of converting the captured solar energy into electricity. The hot-side temperature however, is limited by properties of the working medium. The cold-side temperature is largely determined by the cooling method and the environment available to dissipate the exhaust heat. An indirect but major advantage of this system is that the energy storage is interwoven in the system at no extra cost. It resides in the latent heat of phase change at a high temperature of around 1,000 K. The system can store thermal energy for hours with no electrical performance degradation, or longer with some degradation. This feature makes this technology capable of meeting peak power demands with no added mass or cost of separate energy storage. It eliminates the battery requirement altogether.

Although the solar dynamic technology is not yet proven in space, it offers potential advantages in efficiency, weight, scalability, and the overall cost in high-power spacecraft. The cost advantage comes from the elimination of costly semiconductor PV cells. Such a system can be cost effective in a

Fig. 10.24 RTG I–V–P characteristics of RTGs at various operating age in years

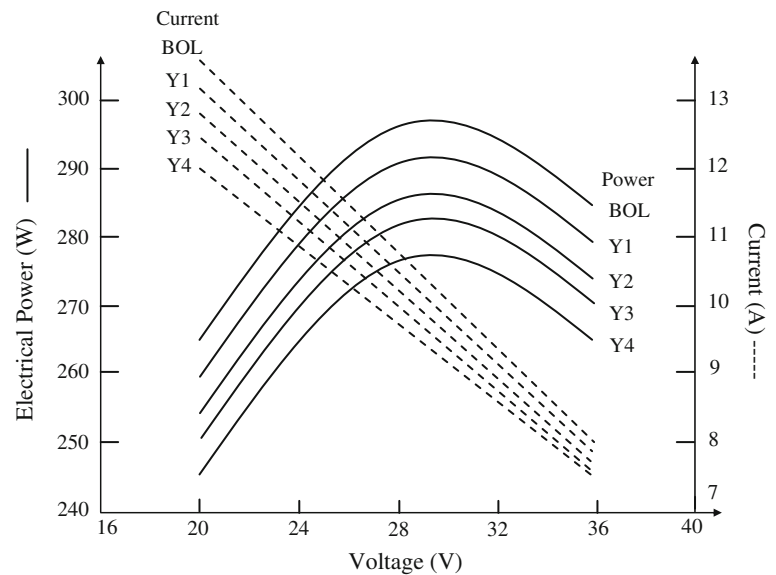
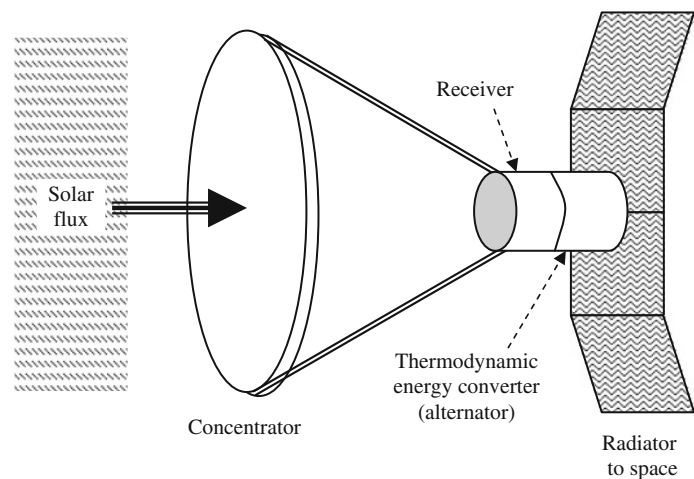


Fig. 10.25 Solar concentrator with dynamic alternator



few kilowatts to hundreds of kilowatts power range. The concept is sufficiently developed for space use in the near future, particularly in high-power LEO missions. It may also find applications in high-power defense spacecraft where large solar arrays can make the vehicle non-maneuverable and vulnerable to the detection and attack by enemy. It has been considered in the past for a 300 kW space station and for a dynamic isotopes power system (DIPS) for space defense.

The efficiency advantage in the dynamic system comes from the higher efficiency of the engine (25–40 %) as compared to silicon solar cells (15–25 %), and higher efficiency of thermal energy storage of the receiver (85–90 %) as compared to the battery efficiency (70–75 %). The greatly improved overall system efficiency as compared to the PV system translates into less solar collection area. This results in reduced drag and relaxed concern regarding station dynamics, approach corridors, and experiment viewing angles. The reduced drag is particularly important because it allows lower

flight altitudes within given constraints of drag-makeup fuel and orbit decay time. At power levels near 100 kW, such as for space-based radar, the PV solar array collector area becomes prohibitive. The solar dynamic power system is expected to find advantageous applications in this power range.

Recent prototype testing of a 2 kW non-optimized solar dynamic systems reported by Mason [14] demonstrated a conversion efficiency of 30 % using 1990s component technologies. Significant improvements in efficiency can be realized for large systems with ratings above 100 kW using newer technology components and optimized design parameters.

10.4.6 Fuel Cell Power

The fuel cell was developed as an intermediate-term power source for space applications. It was first flown on the Gemini V crewed mission in 1965. It has been routinely

used to power NASA's fleet of Space Shuttles that carried components and crew to the International Space Station and other space service missions [15]. The fuel cell resembles a battery in that it converts the chemical energy of a fuel directly into DC electricity. However, unlike a battery, it does not run down in energy and does not have to be recharged. It keeps producing electricity as long as the fuel is supplied. The typical fuel gas is hydrogen or a hydrogen-rich mixture and an oxidant. One pound (450 g) of hydrogen has 52,000 Btu or 15.24 kWh primary energy and requires 8 pounds (<3,600 kg) of oxygen to react.

The fuel cell finds applications in space missions lasting for a few days to a few weeks where the battery is not practical. It also has a potential use as auxiliary power source for orbit transfer vehicles. The regenerative fuel cell integrated with an electrolyzer unit presents an attractive mass saving for LEO satellites requiring large energy storage. It was a candidate in place of the battery for the ISS.

The working of the fuel cell is the reverse of electrolysis. In electrolysis, electricity is injected between two electrodes in water to produce hydrogen and oxygen. In the fuel cell, hydrogen and oxygen are combined to produce electricity and water. The energy conversion is direct from chemical-to-electrical. Since the process is isothermal, the conversion efficiency is not limited by Carnot efficiency. This is unlike chemical-to-thermal-to-mechanical-to-electrical energy converters using steam or an internal combustion engine. It skips the usual combustion step of the conventional thermodynamic power system and converts a high percentage of the fuel's available free chemical energy directly into electricity. The fuel cell efficiency, therefore, can be about twice that of the thermodynamic converter. It is as high as 65 % in some designs, and 75–80 % in solid metal oxide fuel cells developed for ground-base power plants. Its superior reliability with no moving parts is an additional benefit over the thermodynamic power generators.

10.4.6.1 Electrochemistry of Fuel Cell

The fuel cell consists of anode and cathode electrodes separated by a liquid or solid electrolyte. The electrodes are electrically connected through an external load circuit as shown in Fig. 10.26. Hydrogen or a hydrogen-rich mixture is fed to the anode. The hydrogen fuel is combined with oxygen of the oxidant entering from the cathode port. The hydrogen, however, does not burn as in the internal combustion engine. It splits into hydrogen ions (H^+) and electrons (e^-), and produces electricity by an electrochemical reaction. Water and heat are the byproducts of this reaction if the fuel is pure hydrogen. With natural gas (ethanol or methanol) as the source of hydrogen—as in some ground-based fuel cells—the byproducts include carbon dioxide and negligible traces of carbon monoxide, hydrocarbons, and nitrogen oxides.

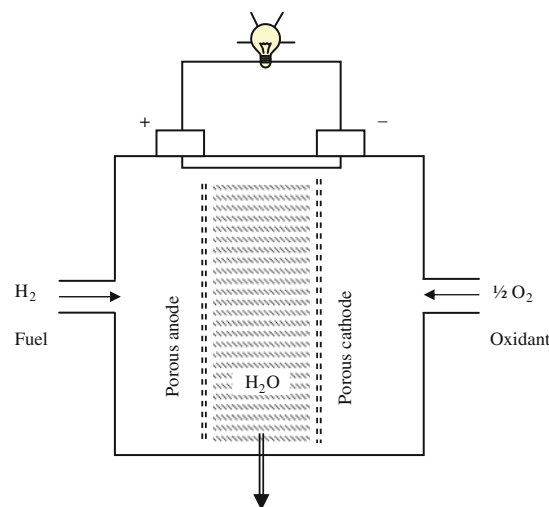


Fig. 10.26 Fuel cell construction and operation

The hydrogen–oxygen fuel cell consumes hydrogen as fuel, oxygen as oxidant, and an aqueous acid solution as electrolyte. Their net reaction is $2H_2 + O_2 = 2H_2O$, and of course energy is released in the process. In one type of fuel cell, the electrons flow from the anode and travel through the external circuit to the cathode, powering the electrical load connected to the terminal. The hydrogen ions migrate through the electrolyte to the cathode, closing the loop. At the cathode, they combine with the oxygen and the incoming electrons from the external circuit to produce water. The kind of ions and the direction in which they migrate varies, depending on the type of electrolyte.

The fuel cell is thus a static electrochemical device that generates electricity by chemical reaction without altering the electrodes or the electrolyte materials. This distinguishes the fuel cell from the electrochemical battery. Unlike the conventional battery, the fuel cell has no electrical energy storage capacity. Hence, it must continuously supply the reactant and withdraw the reaction products during operation.

10.4.6.2 Fuel Cell Performance

The fuel cell works as a voltage source with an internal resistance. The electrical potential appears at the terminals of two electrodes involved in the process. The theoretical value of the fuel cell potential is 1.25 V, which matches that of NiCd and NiH_2 batteries. Multiple fuel cells are stacked in series–parallel combinations using heavy graphite pallets for the required voltage and current, just as the electrochemical cells are in a battery. However, as soon as the current is drawn, the voltage drops significantly due to various losses. Because the primary loss mechanism is ohmic loss in the electrodes, the voltage continues to drop with increasing current. The voltage drop is given by $V_{drop} = (\alpha + \beta \ln J)$, where J is current density at the

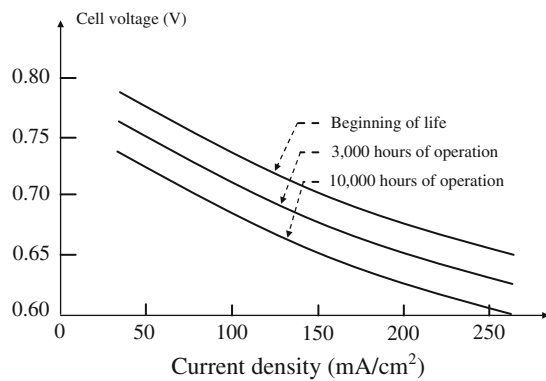


Fig. 10.27 Fuel cell output voltage versus current density and operating hours

electrode surface, and α and β are constants which depend on temperature and the electrode surface.

The theoretical potential difference of 1.25 V between the anode and cathode in the hydrogen–oxygen fuel cell is determined by the difference of the free energy of the reaction product and the fuel and oxidant. This potential is different in different fuel cells depending on the reactions involved. The electrical performance of a fuel cell is represented by the electrode voltage versus surface current density, commonly known as the polarization curve or V–I curve. Ideally, a single $\text{H}_2\text{--O}_2$ fuel cell could produce 1.25 V DC at ambient conditions. Undesirable ions and products of the intermediate irreversible reactions decrease the cell potential, even at open circuit. Further voltage drop under load results from various irreversible polarizations in the cell. The net result of these polarizations is that the practical fuel cell produces between 0.5 and 1.0 V DC at currents of 100–400 mA/cm^2 of cell area. Fuel cell performance can be increased by increasing the cell temperature and reactant partial pressure. A trade-off exists between achieving higher performance by operating at higher temperature or pressure and confronting the materials and hardware problems imposed at more severe conditions.

The practical operating range of the fuel cell is controlled by ohmic loss. The V–I characteristic in this region is very similar to that of a battery, except that the average discharge voltage is lower. The voltage drops approximately linearly with increasing current and also with time, as shown in Fig. 10.27 [16]. At any given time, the terminal V–I relationship can be expressed as $V = V_0 - kI$, where V_0 is the open circuit voltage and k is a constant. The value of k increases and V_0 decreases with time. The power at any operating point is given by $P = VI = [(V_0 - kI)(V_0 - V)/k]$. The maximum power is when $dP/dt = 0$, which occurs at $V = 1/2V_0$, leading to

$$P_{max} = \frac{V_0^2}{4k}. \quad (10.11)$$

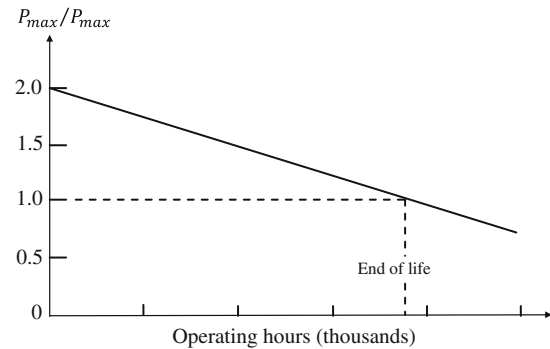


Fig. 10.28 P_{max}/P_{max} ratio versus time in operating hours determines fuel cell life

Table 10.8 Comparison of the performance of various fuel cell

Fuel cell technology	Specific power (W/kg)	Life in hours
Alkaline	100–150	~50,000
Solid polymer	100–150	~50,000
Alkaline (space shuttle)	300–400	3,000–5,000
Lightweight cell under development	600–700	TBD

Unlike the PV cell, the fuel cell does not work in *use the input energy or lose it* mode. It uses the on-board fuel to generate power. For this reason, the fuel cell is not operated at P_{max} until it approaches the end of life. It is rather operated at the maximum fuel efficiency until the EOL.

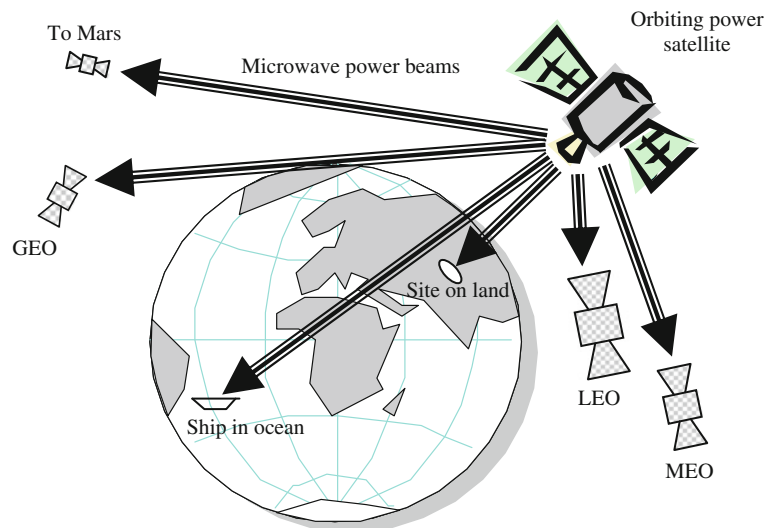
As V_0 degrades with time, so does P_{max} . The open circuit voltage can be expressed as a function of time as $V_0(t) = V_0(0) - K_0h$, where h is the hours since the fuel cell was placed in operation. With a voltage regulating converter between the fuel cell and the load, the life of the fuel cell can be defined as the time it takes for the voltage to decay below the required input voltage, for P_{max} to fall below the required output power. It can be predicted from the V_0 versus time relation. The expected life of the fuel cell is determined as shown in Fig. 10.28 [16].

The performance of various types of fuel cell for space applications is compared in Table 10.8. Compared to a PV array, the fuel cell gives much higher power per kilogram. Flexibility is another major advantage, as it does not need Sun pointing and provides the same power during both day and night. Its disadvantage is that it needs to carry fuel on-board.

10.4.7 Beam Power Satellite

The traditional power system in space uses a solar array and a battery for each satellite. The concept of using a centralized large power-satellite (powersat) has been receiving

Fig. 10.29 Power transmission from central power satellite to multiple satellites



seed funding at NASA for some time. In this concept, depicted in Fig. 10.29, a high-power satellite generates bulk power, which is then transmitted to multiple satellites by laser or microwave beams. The electric propulsion of a spacecraft using beam power from a powersat is also possible. The beam power can also be used for rescue operation in remote regions of Earth, or anywhere in the ocean. The advantages of a central power satellite transmitting beam power to a number of user spacecraft are

- Large PV arrays are replaced by a much smaller power-receiving antenna in the user spacecraft.
- A significantly smaller battery can replace larger batteries as proper orientation with the power satellite(s) will avoid satellite-to-satellite eclipses.
- Longer mission life since the spacecraft life is not limited by its solar array or battery.
- Provides large peak power if and when needed, such as for electric propulsion, thus significantly decreasing the on-board fuel mass.
- Orienting the spacecraft to the powersat beam is much simpler than to orienting to the Sun.

10.4.7.1 Microwave Beam

The power transmission efficiency in space by a microwave beam is given by Lineberry and Chapman [17]

$$\eta = 1 - e^{-\frac{A_r A_t}{d^2 \lambda^2}} \quad (10.12)$$

where, η is the fraction of the transmitted power captured by the receiving antenna, A_r , A_t is the transmitter and receiver antenna area, respectively, d is the distance between the transmitter and receiver, and λ is the wavelength of the microwave power beam. The distance between the powersat and the user satellite may vary. The microwave transmission may be effectively used over short distances.

The frequencies considered for such systems are from 3 to 300 GHz [18] with the corresponding wavelengths from 100 down to 1 mm.

In 2003, an experiment was planned using the Cosmos-1 solar sail, at 800 km altitude, to transmit power by a microwave beam from the Goldstone 100 m antenna. The Goldstone steerable dish radiates up to 1/2 MW power; it was estimated the sail would receive only 1,700 W. The resulting microwave pressure was estimated to accelerate the sail by 10^{-7} g and demonstrate the principle of beaming power to a spacecraft. The acceleration of the sail would have depended only on power and not on the frequency of the beam, however the experiment was not conducted due a failure of the Cosmos-1 launch vehicle. In another experiment at JPL and the University of California at Irvine, a 10 kW, 7-GHz microwave beam in vacuum chamber produced a power density of 1 kW/cm^2 to heat a sail surface to 2,000 K.

Microwave transmitters have been under developments much longer than lasers. They are far more efficient and cost much less. They do not damage the receiving surface as lasers can, and do not refract while passing through air. However, microwaves require much larger antennas for the same focusing ability.

10.4.7.2 Laser Beam

Laser beams may be more efficient over long distances in the 10,000–50,000 km range [19]. The laser system consist of three parts

- First the solar energy collected on the powersat is converted into laser beam using a solid-state solar pumped laser. It consists of a crystal placed in the focus of a parabolic solar concentrator.
- The monochromatic laser radiation is then focused into a beam using an optical mirror of appropriate size. The wavelength of such a laser is equal to $1.06 \mu\text{m}$, and the

conversion efficiency 20–25 %. A focused power laser beam then transmits the power to the user satellite or a rescue site on the Earth at 80–90 % efficiency.

- Finally, the received energy is converted into electrical power using the conventional PV cell. The conversion efficiency of a PV cell under laser illumination is around 50 %. The maximum electrical power output of the silicon PV cell using such a system can be 3,000 W/m² from laser radiation compared to 300 W/m² from natural solar radiation. The solar array requirement on the user satellite is thus greatly reduced, and the battery can be eliminated by avoiding powersat-to-user satellite eclipses.

A powersat, once developed and built, could make the user satellite much lighter and less expensive, so that many could be launched at lower cost per launch. This could open up entirely new kinds of missions in Earth orbit and in interplanetary space at much less incremental cost. The concept is like the 19th century railroad system. Once the tracks are laid, the train itself is a small added expense.

10.4.7.3 Space-to-Ground Power

Driven by the environmental and depletion concerns of the fossil energy sources, the U.S. Department of Energy, NASA and ESA have funded space-based generation of power for ground use. The equivalent mass density of air on Earth with typical moisture and pollution is 1.5 times that of pure air. The solar radiation on a normal Earth surface with air mass 1.5 (AM 1.5) is about 1 kW/m². With 12-hour nights on average, seasonal Sun variations, and overcasts, the annual average energy yield of a ground-based PV system may be around 6 kWh/m² per day. A 400 MW power station on the ground using 20 % efficient PV cells would need 40 million square meters of PV array. Storing sufficient energy to last 5 days without Sun would require 60,000 MWh of energy storage assuming 80 % round trip energy efficiency. A space-based system can reduce the collector area to 1/4th and eliminate the need of energy storage by making the beam power dispatchable on demand.

The performance of a solar array placed on the Earth's surface versus in LEO and GEO is summarized in Table 10.9. It shows that the energy collection per square meter in space is several times higher than that on Earth. Furthermore, it is 50 % higher in GEO than in LEO. One concept study has considered a medium Earth orbit at around 10,000 km altitude for such a powersat for ground use.

Ambitious proposals have been explored for harnessing solar energy for terrestrial use by deploying solar powersats in GEO that could generate power 24 h a day, 365 days a year. One concept study has shown that one satellite with a 146 km² solar array could deliver power equivalent to 10 nuclear power plants on Earth. The 1 km² antenna would transmit power in a sharp 2.4 GHz microwave beam to an Earth receiving station. Here, it would be converted into DC

Table 10.9 Performance of solar array placed on ground and in LEO and GEO

Solar array location	Earth	LEO	GEO
Air mass	1.5 with average moisture and pollution	0	0
Solar radiation (W/m ²)	1,000 in full Sun 500 in partial Sun	1,350	1,350
Incident energy (kWh/m ² per year)	1,643	7,884	11,826
Useful sunlight (h/day)	6 on average	16	24
Launch and maintenance cost in orbit	0	Medium	High

and inverted into 10,000 MW 60 Hz AC and delivered to the distribution system of the electrical power utility. The basic converter would use inductors (perhaps superconducting) to store the energy and boost the voltage.

In the proposed concept, the beamed power is converted to 60 or 50 Hz utility power using high voltage converters. Massive series-parallel connections of numerous converters would be needed to beam gigawatts of power to Earth from space [20]. The space-to-space power transmission would be at low level in W/m², while the space-to-ground power transmission would be in hundreds of W/m². However, a U.S. government regulation limits the microwave radiation to 110 W/m². Therefore, the hundreds of W/m² beam intensity would require a regulation change or special permission to beam to a remote location from where it would be transmitted to populated load centers. This high beam intensity is still a several times lower than the natural Sun intensity of 1,000 W/m² on the ground. Whether it would pose a cancer risk to humans has yet to be resolved.

Several innovative concepts are being studied for collecting solar energy in space and transmitting microwave beams to other spacecraft that may be orbiting the Earth, on an interplanetary mission, or on a planetary surface. NASA's Solar Space Power Exploratory Research and Technology program is investigating systems at power levels ranging from 100 kW to 1,200 MW. The building blocks of such a system are

- A large Sun-oriented solar array that tracks the Sun and generates power at high voltage in the 400–1,000 V range.
- A rotating microwave transmitter in space that tracks a receiving antenna (rectenna)/rectifier station on Earth.
- A microwave beam at several GHz frequency, using solid-state power converters, magnetrons, or klystrons.
- A rotary joint between the solar collector and the transmitter.

- High voltage cables.
- DC to microwave power converters in space and microwave to DC power converters and DC to 50 or 60 Hz AC inverters on the ground.

A concept study for delivering 1,200 MW from GEO to the ground grid has developed the following estimates [21]

- A solar array with 4–5 concentration ratio with futuristic conversion efficiency of 39 % generating at specific power of 1,000 W/kg and 550 W/m² at 1,000 V.
- A 5.8 GHz microwave beam of Gaussian power density distribution with 10 dB taper from the transmitter center to the edge.
- Solid-state transmitters operating at 80 V and the solar array at 1,000 V.
- Power distribution from solar array at 100 kV, so that converters are needed at both ends of the distribution lines. These converters substantially add into the system mass. Using 6,000 V magnetrons and reducing the distribution voltage to 6,000 V could eliminate the transmission voltage converters.
- A rectenna size on Earth about 7,450 m in diameter.
- Overall wireless power transmitting efficiency around 35 %.
- A system for 1,200 MW power to the grid requires a solar array of 7,300,000 m² area, equating to a 2,700 × 2,700 m² or 3 km diameter array in space. The rectenna on the ground is estimated to be 44,000,000 m² area or 7.5 km diameter. The ground receiver would be about six times the solar array area because of (a) power losses in various components, and (b) the beam power density being limited under the federal regulation on microwave power.
- The mass of the above concept satellite is estimated to be 22,500–30,000 metric tons at launch and 17,000–22,000 metric tons in orbit. A great many technology developments and demonstration are needed to make the cost per kWh delivered to the ground competitive with the conventional ground-based grid power [22–24].

Acknowledgments Several figures and tables in this chapter are reproduced with permission from Patel, M. R., “Spacecraft Power Systems”, CRC Press, 2005. The author is also grateful to two industry experts for reviewing the manuscripts and providing valuable feedbacks. They are Mr. James E. Haines, retired head of power system group of European Space Agency, and Mr. Abbas Salim, retired senior staff engineer of Lockheed Martin Corporation.

References

- Hyder, A. K. et al., “Spacecraft Power Technologies,” Imperial College Press/World Scientific Publishing, London, 2003.
- Marshall, C. G. et al., “Example of a prototype lightweight solar array and the three promising technologies it incorporates,” *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 01-2550.
- Frohlich, R. C., “Contemporary measures of the solar constant: The solar output and its variations,” Colorado Associated University Press, Boulder, CO, 1977, pp. 93-109.
- Green, M.A., Emery, K., Hishikawa, Y., Warta, W., Dunlop, E.D. 2011. Solar cell efficiency tables (Version 38), *Progress in Photovoltaics: Research and Applications* 19, 565-572.
- Parez, M. E. et al., “Energy storage for space applications,” *Proceedings of the 36th Intersociety Energy Conversion Engineering Conference*, ASME, 2001, pp. 85-89.
- Hojnicki, J. S. et al., “Space Station Freedom Electrical Performance Model,” NASA Glenn Research Center, Report No. TM-106395, 1993.
- Haines, J. E., “Inner Planets sample return missions, the challenge for power systems,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 2483.
- Brandhorst, Jr, H. W. and Chen, Z., “PV approaches for near-Sun missions,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 2631.
- Choi, M. K., “Power and thermal systems with thermoelectric generators at 930°C for solar probe inside 0.1 au,” *Proceedings of the 36th Intersociety Energy Conversion Engineering Conference*, ASME, 2001, Vol. II, pp. 1161-63.
- Elbuluk, M. E. et al., “Low temperature performance evaluation of battery management technologies,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No.01-2543.
- Nagasubramanian, G., “Low temperature electrical performance characteristic of Li-Ion cells,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 01-2462.
- Croft, H., Staniewicz, R., Smart, M. C., and Ratnakumar, B. V., “Cycling and low temperature performance operation of Li-ion cells,” *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference*, AIAA, 2000, Paper No. 27-AP-B1.
- Smart, M. C, Huang, C. K., Ratnakumar, B. V., Surampudi, S., and Sakamoto, J. S., “Factors affecting Li-Ion cell performance,” *Proceedings of the 37th Power Sources Conference*, Paper No. 239, 1996.
- Mason, L. S., “A solar dynamic power option for space solar power,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 01-2601.
- Oman, H., “Fuel cells power for aerospace vehicles,” *IEEE Aerospace and Electronics System Magazine*, Vol. 17, No. 2, February 2002, pp. 35-41.
- Babasaki, T., Take, T., and Yamashita, T., “Diagnosis of fuel cell deterioration using fuel cell current-voltage characteristics,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 01-2575.
- Lineberry, J. T. and Chapman, J. N., “MHD Augmentation of rocket engine for space propulsion,” *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference*, AIAA, 2000, Paper No. 3056.
- Koert, P. and Cha, J. T., “Millimeter Wave Technology for Spec Power Beaming,” *IEEE Transactions on Microwave Theory and Technology*, Vol. 40, No. 6, June 1992, pp. 1251-58.
- Grechnev, A. B. et al., “Centralized power as basis of new philosophy of space power engineering,” *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference*, SAE, 1999, Paper No. 2436.
- Kusic, G., “Conversion of beamed microwave power,” *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference*, AIAA, 2000, Paper No. 3071.
- SAIC and Futron Corporation, “Space Solar Power Concept Definition Study,” NASA Report No. SAIC-99/1016, February 1999.

22. Mankins, J. C. and Howell, J., "Overview of the space solar power exploratory research and technology program," *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference, AIAA, 2000*, Paper No. 3060.
23. Carrington, C. et al., "The abacus/reflector and integrated symmetrical concentrator concept for space power collection and transmission," *Proceedings of the 35th Intersociety Energy Conversion Engineering Conference, AIAA, 2000*, Paper No. 3067.
24. Lynch, T. H., "Sun tower PMAD architecture," *Proceedings of the 34th Intersociety Energy Conversion Engineering Conference, SAE, 1999*, Paper No. 2441.

Further Reading

25. Patel, M.R., "Spacecraft Power Systems", CRC Press, Boca Raton, 2005.
26. Hyder, A. K. et al., "Spacecraft Power Technologies," Imperial College Press/World Scientific Publishing, London, 2003.

Claudio Bruno

11.1 Fundamentals of Rocket Propulsion in Vacuo

Propulsion, from the Latin words *pro* meaning before or forwards and *pellere* meaning to drive is, as this Latin word implies, the art of pushing ahead, and in space this means to push a spacecraft to accelerate it.

In space, with no solid or fluid available, Newton's third law of motion tells that the push may be produced only by taking part of a spacecraft mass, the mass of the propellants, m , and expelling it at a speed v_e . When pushing out matter, the inertia of the expelled mass produces an equal and opposite push or 'thrust' against the rest of the spacecraft and the spacecraft is accelerated.

On Earth, the push to accelerate a body may be against the ground, as in walking, when the acceleration is the result of the pavement pushing against the shoe soles or foot and a push force is applied to the soil by friction. In swimming, or rowing, a force (pressure) is applied to water by using a solid surface, oar or arm, and the resulting pressure accelerates the oar or arm in return. In space, there is nothing to push 'against', and it is the action of ejecting mass from a spacecraft, that is, of imparting a momentum to the mass ejected, that applies the exactly opposite reaction, resulting in a change of the momentum of the spacecraft. In general, the mass ejected may be already in motion with respect to the spacecraft, so the thrust force F is therefore described as the difference between the momentum inside the engine of the mass to be ejected, and that when leaving the spacecraft. The scalar form of F is

$$F = \frac{dm}{dt} v_e \quad (11.1)$$

and is generally a function of time. In propulsion systems where the exhaust is obtained by a thermodynamic expansion and the mass is gaseous, there is an extra term $A_e(p_e - p_a)$ contributing to the total F , see Sect. 11.9, due to the fact that there is a difference of pressure between the engine exit and the ambient environment. In space, the ambient pressure is of course practically zero. The total impulse, I_{tot} , corresponding to a thrust force F acting for a total time t , is defined as

$$I_{tot} = \int_0^t F(t) dt. \quad (11.2)$$

The total impulse is a measure of the total change of momentum available from a propulsion system, and is especially important when the thrust changes over time. Examples are solid propellant rocket boosters, where the thrust is continuous but made to vary over time with a specified law, and attitude control rockets that must deliver short thrust pulses many hundreds of times.

As introduced in Chap. 4, the rate at which mass is ejected and its exhaust velocity, v_e , determines the thrust and specific impulse, I_{sp} , which is the ratio of thrust to Earth-surface weight flow rate, $\dot{m}g$, where \dot{m} is the propellant mass flow rate

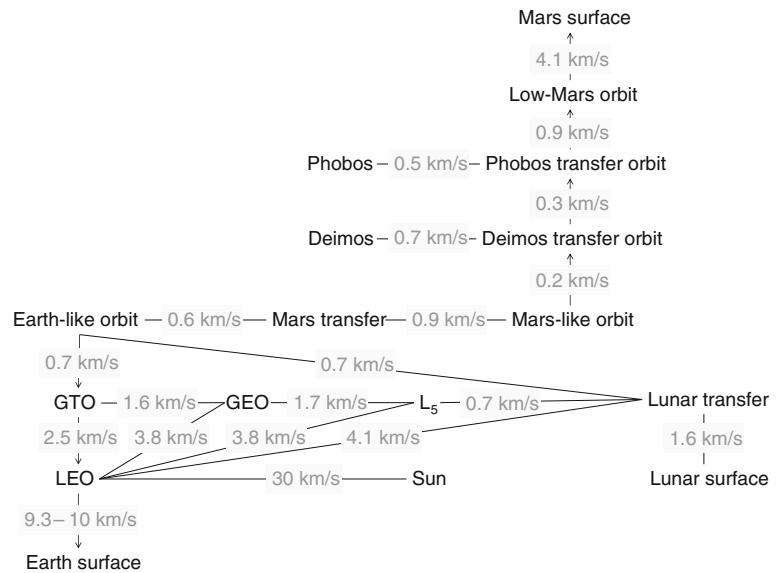
$$I_{sp} = F/\dot{m}g = v_e/g. \quad (11.3)$$

Note that this equation is the same as Eq. 8.2, and was introduced in Eq. 4.146. The units of I_{sp} are seconds, where the amount of propellant is measured as a weight, or force, such as kilogram-force or kiloponds.¹ However, if the amount of propellant is measured as a mass then specific impulse has units of velocity. The conversion constant between the two versions of specific impulse is g , standard gravity defined as precisely 9.80665 m/s².

C. Bruno (✉)
Thermal and Fluid Sciences, United Technologies Research
Center, 411 Silver Lane, MS 129-29, East Hartford, CT 06118-
1127, USA
e-mail: BrunoC@utrc.utc.com

¹ A kilopond, kp, (or kilogram-force, kgf) equal to the magnitude of the force exerted by one kilogram of mass in a standard gravity field (9.80665 m/s²). 1 kp = 9.80665 N. Kilogram-force, or kilopond, is a non-standard unit and does not comply with the SI Metric System but remains widely used.

Fig. 11.1 Minimum ΔV for Near-Earth and Mars missions with Hohmann transfers; arrows denote the possibility of using an aerobraking maneuver. *Image* Malcolm Macdonald



From Eq. 11.3, it is seen that where the amount of propellant is measured as a mass, the specific impulse is the ideal velocity of the ejected mass. In an ideal, one-dimensional acceleration process where no energy is wasted, the velocity of the momentum and the specific impulse coincide. In a liquid rocket engine (LRE) ideal v_e and (gI_{sp}) do not coincide as the gas expands only to the pressure at the nozzle exit, which is never zero, and hence some of the kinetic energy of the mass ejected does not do useful work and is instead wasted; the thermodynamic efficiency of the rocket engine cycle is never 100 %.

Using the specific impulse, and as introduced in Eq. 4.146, the ideal rocket equation, also known as Tsiolkovsky's rocket equation, gives the change in mass of the spacecraft due to the expelled mass for a given change in velocity, ΔV , as

$$m_f = m_0 \exp(-\Delta V/gI_{sp}) \quad (11.4)$$

where m_0 is the initial mass of the spacecraft and m_p is the propellant mass, such that the final mass is $m_f = (m_0 - m_p)$. Figure 11.1 shows the practical, rather than ideal, ΔV for missions in the vicinity of the Earth and Mars.

Inefficiencies and perturbations such as finite burn losses, as discussed in Chap. 4, or thrust misalignments add to the total impulse and the ideal ΔV required to achieve a given orbit, and thus add to the total mass consumption. Reaching low Earth orbit (LEO) requires an ideal ΔV of about 8 km/s (9.3–10 km/s taking into account gravity losses and aerodynamic drag), so to reach a reasonably high mass ratio, Eq. 11.4 shows that specific impulse should be of the same order, because the higher the specific impulse the lower the mass consumption.

Unfortunately, chemical energy propulsion is fundamentally incapable of generating specific impulses higher than 5,000–5,500 m/s, which means that simply to reach

LEO the mass ratio of the payload will be low, even with multi-stage rockets. Only propulsion systems with specific impulses at least of the same order as the ΔV needed by the specific mission can reduce mass consumption. Thus, short of inventing some sort of 'space drive', the price to pay to accelerate or decelerate a spacecraft (to acquire a positive or negative ΔV) is the mass that must be ejected. This mass consists of propellant(s) that must first be lifted to orbit at great expense, currently about seven to twenty thousand US dollars per kilogram, depending on the launch provider. Once orbited, the propellant(s) mass must be accelerated by a propulsion system using some force. The work done by this force is ideally equal to the kinetic energy acquired by the matter that is expelled. These considerations introduce the next topic: What type of force is available in order to apply Newton's third law to space propulsion?

Current understanding of physics ('the Standard Model') shows only three fundamental interactions (also called fundamental or interactive forces). In increasing order of magnitude, they are: gravitational; electro-weak (the result of the unification of electro-dynamic and weak force in the 1980s); and strong, or nuclear, interaction. The gravitational interaction acts over (not 'at') large distances, and its associated particle, the graviton, has been postulated but, so far, not found; see also related discussion in Sect. 4.1.3. Its magnitude becomes observable only with masses of the order of planetary masses. The electro-weak interaction is responsible for Coulomb and Lorentz forces, and thus holds together atoms and molecules ('chemical bonding') and a manifestation of its effect is combustion. Its magnitude is about 10^{17} times that of gravitation. The nuclear interaction keeps together the nuclei of atoms, preventing their disintegration due to Coulomb repulsion. It acts at nuclear distances, about 1 fermi (10^{-14} m, which is, by no coincidence, the size of a

Table 11.1 Energy potentials of the three fundamental interactions, where α is the fraction of mass convertible to kinetic energy based on $E = mc^2$

Interaction	Potential	α	Energy density, J (J/kg)
Gravity	Gravitational	10^{-27}	10^{-11} (two 1-kg masses at 1-m distance)
Electro-weak	Chemical (H ₂ /O ₂ combustion)	1.5×10^{-10}	1.4×10^7
Strong	Nuclear fission (²³⁵ U)	9.1×10^{-4}	8.2×10^{13}
	Fusion (D-T)	3.8×10^{-3}	3.4×10^{14}
	Metastable (^{180m} Ta)	2.0×10^{-7}	1.8×10^{10}
	Annihilation (p ⁺ - p ⁻)	1.0	9.0×10^{16}

nucleon). Its magnitude is about 10^7 times that of Coulomb. To each of the three forces is associated a potential. When the potential decreases, energy is released. Its effect varies with the respective potentials, but is always in the form of kinetic energy. The potential energy density (per unit mass) of the three interaction is quantified in Table 11.1, in which ‘metastable nuclei’ are nuclei of isotopes where the spatial configuration of nucleons is not that corresponding to the minimum energy state, and thus is able to ‘relax’ to that minimum by yielding energy without fissioning. ‘Annihilation’ is the process whereby mass is completely converted into energy following Einstein’s famous $E = mc^2$, that is, $\alpha = 1$.

Table 11.1 shows that, per unit mass, the nuclear potential is the largest. The electro-weak potential is about 10^6 – 10^7 times lower, and the gravitational about 10^{24} times lower. The gravitational potential can be exploited by flyby or gravity assist maneuvers, as discussed in Chap. 4. Such maneuvers are becoming routine, and can save significant propellant mass. For instance, a Pluto mission may direct the spacecraft toward Venus, Mars, or the Earth itself, and the spacecraft will acquire kinetic energy at the expense of the planetary gravitational potentials, and then ‘swing by’ with increased speed. The Cassini probe left Earth orbit at close to 12 km/s, and after repeated gravity assists, reached the Jovian system at more than 50 km/s. These savings in propellants mass are at the cost of much longer trajectories and travel time. Perhaps that is not critical in scientific missions, but it becomes intolerable for crewed missions, where solar and galactic radiation doses to a human crew may be hundreds or thousands of times the standard dose in a year on Earth, which is a fraction of a milli-Sievert (mSv). It has been estimated that a crewed Mars mission lasting of the order of a year will result in a dose to crew of order 1 Sv, with a significant probability of getting some form of cancer.

Propulsion systems exploit one of these three forces, and the vast majority exploit the second. Gravitation is too weak to become the driver of a space engine. The nuclear force was exploited in the 1950s through to the 1970s to build nuclear thermal rockets, where the thermal energy produced in a nuclear reactor (NR) is used to heat a working fluid that

is ejected to produce thrust, see Sect. 11.9. More recently, nuclear reactors have been proposed to power electric rocket engines (electric thrusters, or ET). Because of public diffidence and fear, no nuclear-powered rocket or ET has been [officially] tested by the US since the early 1970s. The obvious advantages of nuclear power (NP) notwithstanding, space propulsion has remained focused mostly on chemical propulsion (CP), where exothermic reactions produce hot gas that by expanding thermodynamically in a nozzle, converts its potential energy (enthalpy) into kinetic energy. More recently, there has been work on electric propulsion (EP), powered by solar panels. EP is based on direct application of electrostatic (Coulomb) or electro-dynamic (Lorentz) forces to ionized chemical species or particles. Electrothermal operation (electrically heating and then expanding propellant) is also used.

The performance of rockets/thrusters that do not operate based on thermodynamic expansion—for instance, all electric thrusters, where an electric or magnetic field accelerates charged particles—depends on the specifics of the thruster. In thermodynamic thrusters the maximum I_{sp} is limited by the energy density available. With chemical propulsion the maximum practical I_{sp} in vacuo is about 4.5 km/s using the LOX/LH₂ combination, equivalent to about 465 s. Specific impulse could exceed 500 s by replacing LOX with liquid fluorine (FLOX), but environmental considerations and logistics forbid its use. With EP, the I_{sp} is limited by technology because the Coulomb or Lorentz forces that can be applied to ions are limited by voltage breakdown and arcing, and by the intensity of the magnetic field that can be produced by a magnet, roughly 5 kV and 8–10 T, respectively. Thus, substantially higher specific impulse is feasible with EP, but at the expense of increasing thrust F , the jet power $= \frac{1}{2}Fv_e = \frac{1}{2}FI_{sp}$. The term ‘ FI_{sp} ’ scales with v_e^3 and grows very rapidly with increasing specific impulse. At fixed electric power the product F times I_{sp} is constant.

Solid and liquid chemical propellant rocket engines (solid rocket engines are traditionally called solid rocket motors, or SRM) are the workhorse space propulsion systems. They are conceptually similar to those that power launchers to orbit,

and materials excepted, their technology is still where it was at the dawn of the space age, between 50 and 60 years ago. Improvements have been in terms of materials and structures, and especially ubiquitous electronics. Specific impulse has inched upwards at the rate of maybe a fraction of a second per year and is now topping at 454 s for space-operated LOX/LH₂ liquid rocket engines, see Sect. 11.3, and at 260–295 s for solid rocket motors, see Sect. 11.2. Hybrid rockets, where the oxidizer is typically a liquid like LOX or nitrous oxide, N₂O, or hydrogen peroxide, H₂O₂, injected inside a hollow solid fuel grain, have been proposed, developed and ground-tested, but even after many years of development are still not as efficient, and are used only for some sounding rockets. Their future may brighten because of the high visibility acquired by application to the SpaceShipOne and SpaceShipTwo suborbital vehicles. The specific impulse of hybrid engines is typically higher than that of solid propellants, but lower than liquid rocket engines. These engines are briefly discussed in Sect. 11.4.

Electric propulsion is relatively new as an in-space propulsion technology, although its roots go back to the work done by Ernst Stuhlinger (1913–2008) in Germany during World War II (Stuhlinger was a member of the von Braun team). In most electric propulsion applications, the electro-weak force, Coulomb or Lorentz, is applied directly, not in the form of the rearrangement of chemical bonds. The propellant (Hg, Cs, Ar, Xe, H₂, Li, et cetera) must first be ionized, requiring between 3.89 and 15.5 eV of energy. In electrostatic thrusters it is the Coulomb force that accelerates ions, hence their name, ion thrusters, or gridded ion engine (GIE). This is the simplest type of EP. Alternatively, ionized propellant may be injected in a magnetic and electric field manipulated spatially so that it is the Lorentz force that accelerates it. In some electric thrusters, part of the thrust results from applying the Lorentz force, and part from the energetic collisions in the plasma driven by the presence of the electric field. Inelastic plasma collisions convert kinetic energy into heat, and the resulting hot and neutral gas can be expanded as in a conventional nozzle. This second family of electric thrusters includes magnetoplasmadynamic (MPD) thrusters, see Sect. 11.8. The physics of this technology is less advanced than that of gridded ion thrusters, mostly because the magnetohydrodynamics (MHD) plasma is not well understood and instability modes are a commonplace feature. Hall thrusters can be looked upon as either downstream cathode MPD devices or electrode-less electrostatic thrusters.

Some electric thrusters do not require the creation of a plasma. This is the case for resistojets, where the propellant is heated by a resistor, for some ion thrusters that use caesium, where ionization occurs simply by contact with heated tungsten or colloidal fluids and the colloid droplets

are charged electrostatically, and for field emission (FEFP) thrusters where ions are extracted from molten metal. Hydrazine resistojets are used on many geostationary telecommunication satellites.

The ultimate propulsion system is based on the strong, or nuclear, interaction, and was the object of much research and development in the 1950s and 1960s, see Sect. 11.9. In the same section, solar sails are also briefly described.

The subdivision of propulsion technologies based on fundamental physics (‘the Standard Model’) does not cover exotic concepts such as zero-point energy, artificial space curvature, or wormholes in spacetime. Some of these exotic proposals are grounded in current physics (general relativity), but are still immature as far as the method of exploiting them for in-space propulsion. Some are built on suspect physics [1].

11.1.1 Performance

Thermodynamic rockets accelerate and eject a hot gas by thermodynamically expanding it in a nozzle, that is, a duct in which the cross-sectional area increases from the throat A_t to the exit area A_e , see Sects. 11.2–11.4. Truly accurate performance prediction of thrust requires the solution of the reactive, multiphase, time-dependent, three-dimensional Navier–Stokes equations. However, for most (but not all) applications, the flow is mainly one-dimensional, and the solution may be approximated analytically if constant c_p and c_v are assumed, where c_p is the specific heat at constant pressure and c_v is the specific heat at constant volume, if the walls are adiabatic, and if the Reynolds number $\gg 1$. Viscous effects are then limited to the boundary layer close to the walls, and under these assumptions may be, if necessary, calculated separately from the one-dimensional flow. The simple one-dimensional approximation is not only computationally convenient, it is also useful to single out key parameters and understand fundamental physics, see Sect. 11.2.

For propulsion systems that do not use the thermodynamic expansion of a gas, for instance electric thrusters, the performance depends on the force that accelerates charged particles. This force, Coulomb or Lorentz, depends in turn on the specific configuration of the electric or magnetic fields, either applied or self-produced by moving charges; thus, no simple general relationships for thrust and I_{sp} holds for electric thrusters. It is however useful to remember that using energy arguments, the velocity due to electrostatic acceleration (the I_{sp}) scales as the square root of twice the potential, and that due to the Lorentz force this scales with the magnetic induction, \mathbf{B} . In other words, the magnetic pressure scales with B^2 . Thus, the performance of electric thrusters must be found on a case-by-case basis.

11.2 Solid Propellant Rockets

Solid propellant (SP) can be divided into two types: composites and double base. Composites combine separate fuel and oxidizer phases, both in solid form, inside a polymer that is liquid in its original form, and solidifies during a thermochemical process called ‘curing’. This class of solid propellant is almost invariably used in boosters (‘solid boosters’) supplying most of the thrust needed by space launchers and ballistic missiles at liftoff. A second class of solid propellant consists of a solid solution of two energetic materials, each containing molecules consisting of O, C, H and N atoms. Their reaction is exothermic and forms hot CO_2 , H_2O , NO , and N_2 gases. The most important double base solid propellant is the combination of nitrocellulose (which has good mechanical properties and is fuel atom-rich) and nitroglycerin (the more energetic and O-rich base). Additives invariably supplement both classes to improve their manufacturing, mechanical, thermal and combustion properties, and their nature and processing are trade secrets.

Once ignited, the solid propellant releases combustion gas that pressurizes the motor. The burning rate of propellant increases and continues while the pressure stays above the minimum pressure for steady operation (the so called ‘pressure deflagration limit’, or PDL). Below that pressure the SP extinguishes. Thus, there is no way to stop or control solid propellant combustion once started, except by sudden depressurization, for instance bursting the case open with a detonating cord. This is one of the drawbacks of solid propellant propulsion. On the plus side, solid rocket motors are logistically easier to operate and, for the same total impulse, are less costly than equivalent liquid rocket engines.

11.2.1 Solid Rocket Motor; Main Features

This section focuses on composite solid propellant, the most common type for in-space propulsion. Rockets burning solid propellant are also called solid rocket motors, and they consist of a case to host the burning ‘grain’ (see Fig. 11.2) and withstand the internal combustion pressure and temperature, a nozzle to expand the combustion gases, and ignition and thrust vectoring systems. The case may be metallic (for example, steel segments as in the US Shuttle boosters) or be made of a resin in the form of a continuous filament that is wound and cured until rigid (as in the P-80 first stage of the European Vega launcher). Filament winding is the more costly technology, requiring curing in a controlled atmosphere (inside an autoclave), but it results in stronger and lighter cases that enable a larger payload fraction. The parts of the case that have double curvature surfaces (the top and bottom ‘polar bosses’) to seal the cylindrical case are typically, but not always, metallic, and

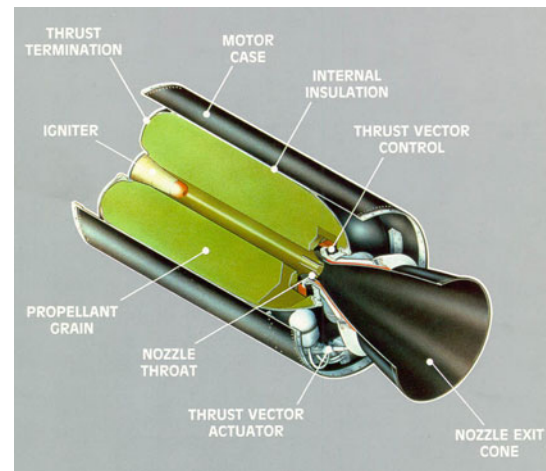


Fig. 11.2 Cut-away view of a nominal SRM for in-space propulsion. The burning surface shown is the internal surface of the cylindrical cavity (the engine ‘port’)

often a titanium alloy. The solid propellant is cast as a ‘grain’, that may consist either of a single cast cylinder of solid propellant, or several cylindrical segments, each manufactured separately for engineering convenience and piled on top of each other prior to launch (this was the case of the Shuttle boosters, for instance). Grain combustion may take place on the flat cylindrical base, as in a cigarette, where the burning surface regresses but its area does not change over time. More commonly, however, the grain is hollow, with the hollowed volume being called the ‘port’, and combustion takes place on, and consumes, the exposed surface of the propellant so that the area may change as combustion progresses. The instantaneous shape and area of the port determines the rate of gas production and thus the instantaneous thrust. Multi-segment solid propellant grains are rarely used for in-space propulsion because the thrust and total impulse, I_{tot} , requirements are much less than for launchers, and therefore these SRM are single-grain. Common applications are orbit raising (LEO to GEO), orbital changes and insertion into final orbits, maneuvering, and stage separation. More complex or repetitive maneuvers may need real-time precise control of thrust and burn time, and so are performed by liquid rocket engines. It should be noted that the use of SRM for in-space propulsion is reducing due to the generation of slag particles that can cause a debris risk to other spacecraft.

Combustion in a SRM converts solid propellant into high temperature gases. Ideally only the port surface burns. Any accidental cracks increase the burning surface and result in excessive pressure and possible SRM destruction. Great care must be taken in making sure that ignition and combustion take place only on the exposed port surface, and this is accomplished by insulating or making inert all other surfaces that might accidentally come into contact with hot burnt gas

and ignite. Sealing these surfaces with ablatives to prevent them from igniting is a key manufacturing step. As in liquid rocket engines, SRM pressure is the result of the balance between mass flow of gases produced during combustion, a function of the burning rate that depends on pressure, and the mass flow of gases that the nozzle can eject, also a function of pressure. During steady-state operation the two must be equal, and that fixes the nominal pressure in the port of a SRM, which can reach 90–150 bars.

A generic SRM for in-space propulsion is shown in Fig. 11.2. The internal insulation (some type of rubber or polymer) shields the case from excessive temperatures and inhibits ignition. The nozzle is shaped as a simple conical bell and may be metallic or high performance resin, while its inner wall is ablative because no regenerative cooling is possible with solid propellant. This wall ablates endothermically, and if its thermal conductivity is sufficiently low it will keep the nozzle at an acceptable temperature. The igniter, or initiator, is a self-contained piece of hardware. It injects hot gas, or hot sparks, formed and released either by a miniature rocket or by a pyrophoric ‘squib’, into the engine port. Either one will ignite the surface at many places simultaneously. This is important to prevent asymmetrical or dramatic thrust spikes during the transient and the formation of uneven burning surface. For large SRM the ignition system may be a chain of devices. The first (a ‘squib’) may be pyrophoric, and ignites the fuel in a bigger chamber containing more energetic material (here pellets of solid propellant mixed with a very reactive metal, say boron). Combustion in this second chamber produces a flame jet that spreads combustion over the entire surface of the grain port.

Note in Fig. 11.2 the gimbal nozzle to control the direction of the thrust and its actuator subsystem. Due to combustion pressure (up to about 150 bars) the gimbal system must be gas-tight, on pain of catastrophic failure. This is obtained by a complex system of flex-joints between the bottom polar boss and the nozzle, one of the proprietary technologies in most SRM. Alternatively, the SRM may have a fixed nozzle flanged to the polar boss, which is more typical of orbit raising SRM.

Once the igniter starts the SRM burning, the grain surface starts to regress. Short of depressurizing the case by an explosive device, thrust cannot be stopped, and is controlled only by the change over time of the surface shape. Each mission requires a thrust profile, so the shape of the burning surface must be precisely designed in advance. This is unlike liquid rocket engines in which valves regulate the flow of fuel and oxidizer and the thrust can be varied; in some specialized types of liquid rocket engines from 10 to 110 % of nominal, but 95–110 % is far more common. The lack of real-time thrust control is perhaps the single greatest disadvantage of SRM compared to liquid rocket engines.

Their advantage is the simplicity of operation, requiring no tanks to be filled. That is why all modern military ballistic and tactical missiles are powered by SRM.

11.2.2 Solid Propellants

Ideally, a solid propellant should have high energy per unit mass, produce gas of low molecular weight (MW), be dense, be chemically and physically stable (e.g. with respect to changes in temperature or humidity), be safe to handle and not accidentally ignited by static electricity, impact, friction or high ambient temperature, be mechanically strong, once cured, and have low thermal expansion coefficient. It should be easy to pour uncured in order to cast in a mold, and should not produce excessive smoke when burning. Recently, low environmental impact has also become an important feature. Ideally, it should also be inexpensive. However, that is not the case with solid propellant. The cost of the standard Ariane 5 booster propellant (hydroxy-terminated poly-butadiene, HTPB 14–18, a composite with 18 % aluminum content and 14 % binder) is between 50 and 75 €/kg, while the throat inserts, which are made of carbon–carbon, cost from 1,000 to 1,600 €/kg depending on the fiber structure. This cost is higher than, say, the liquid hydrogen/oxygen mixture used for the Ariane Vulcan or Shuttle main engines of about 3 €/kg ($LH_2 \sim 1$ €/liter, $LOX \sim 0.2$ €/liter). Nevertheless, the SRM total cost is lower than that of a cryogenic stage, on a total impulse basis.

As previously stated, there are two families of solid propellant: double base and composites. Double base solid propellants are made by partially dissolving nitrocellulose (NC) a solid using nitroglycerine (NG), a liquid explosive that is very sensitive to shocks. Both molecules contain C, N, H and O atoms in reasonable proportions, enabling complete combustion to CO_2 , NO , NO_2 and H_2O . After careful mixing of the two, the resulting solution is solid and much less shock-prone. These solid propellants can be worked and shaped in many ways, but must be made opaque to light, and in particular to infrared light, in order to prevent accidental ignition by radiation transmitted from the burning surface to within the bulk volume. This is done by adding carbon black to the NC/NG solution. Historically, double base propellants were introduced a century ago as ‘smokeless powder’ to replace traditional black powder in all types of munitions, from handguns to large caliber guns. Their cost is moderate, compared to that of composites. In space applications only composites are used, as double base propellants have insufficient mechanical strength to enable them to be shaped as large ‘grains’.

Composite solid propellants are a heterogeneous mix of an oxidizer and a fuel, in a matrix, or binder, that encases and supports them mechanically, plus additives to improve

mechanical, thermal or performance characteristics. To obtain complete combustion, the stoichiometry of these heterogeneous components must be controlled during manufacturing. Fuel and oxidizers are finely ground so they form a solid suspension in the liquid polymeric matrix to be poured in the die and then cured. Carboxyl-terminated polybutadiene (PB) and hydroxyl-terminated PB are the polymers of choice in current solid propellant applications. Most frequently, the oxidizers are perchlorates or nitrates, containing chlorine dioxide or nitrogen oxide groups. They are bound to ammonium ions (NH_4), potassium (K) and Sodium (Na) groups, respectively, forming oxidizing salts that are commercially available. Ammonium perchlorate (AP) is the most common oxidizer, constituting 60–86 % of the solid propellant. The size spectrum of the oxidizer is broader than that of the fuel, and may go from 10 to 400 μm . This facilitates the right ‘packing’ of the oxidizer and fuel, and ensures the correct stoichiometry. Currently used fuels are aluminum (Al) powders, with tailored size distribution. Boron (B) zirconium (Zr) and magnesium (Mg) have been proposed or tested instead of Al. All metals produce combustion temperatures in the 2,800–3,500 K range, and their presence as a solid or liquid phase tend to dampen combustion instability. The concentration of aluminum in composites is 5–20 %, with 16–18 % more common. The polymeric binder constitutes 12–15 % of the composite, and competes with the aluminum for the oxygen in the AP (ammonium perchlorate) or AN (ammonium nitrate) salt.

Additives ‘cure’ and bond the mixture during the casting process, while catalyzers reduce or increase the burning rate and plasticizers facilitate pouring and casting. A partial list of chemicals forming a composite solid propellant is given in Table 11.2. Catalysts increase or reduce the burning rate. Some do this in certain ranges of pressure, for instance producing a nearly constant burning rate (thickness of solid propellant consumed per unit time, in mm/s or cm/s) between 30 and 60 bars (‘plateauing’ catalyzers).

A high performance, high specific impulse composite contains about 70 % AP, 20 % Al and 10 % binder. The densities of these three phases are 1,950, 2,700 and 900 kg/m^3 , respectively. The average density is about 1,800 kg/m^3 for this Al-rich formulation.

Combustion of aluminized AP-HTPB propellants forms liquid Al_2O_3 , CO and CO_2 , H_2O , HCl and other products in much smaller quantity; the alumina formed by burning Al is liquid at the combustion temperature in the SRM (2,900–3,200 K) but becomes solid at 2,300 K, eroding mechanically and depositing as a whitish coating in the nozzle. To maintain the nominal nozzle area (and thrust), carbon–carbon inserts that form a harder ‘collar’ are applied to the nozzle throat.

Table 11.2 A partial list of chemicals forming a composite solid propellant

<i>Oxidizers</i>	AP, ammonium perchlorate
	AN, ammonium nitrate
	NP, nitronium (NO^+) perchlorate
	KP, potassium perchlorate
	RDX, cyclo-tri-methylene tri-nitramine
	HMX, cyclo-tri-methylene tetra-nitramine
<i>Metal Fuels</i>	Al, Aluminum
	Mg, Magnesium
	Be, Beryllium
	B, Boron
	Zr, Zirconium
<i>Matrix/Binder</i>	CTPB, Carboxy-Terminated Poly-Butadiene,
	HTPB; hydroxy-terminated poly-butadiene
	PS, polysulfide
	PVC, polyvinyl-chloride
	PU, poly-urethane.
<i>Curing agents</i>	MAPO (tris (1-2-methyl) aziridinyll phosphine oxide
	IPDI, iso-phorone di-isocyanate
<i>Bonding agents</i>	MAPO
	TEA, tri-ethanolamine
<i>Plasticizers</i>	DOA, di-octyl adipate
	IDP, iso-decyl pelargonate
	DOP, di-octyl phtalate
<i>Burn rate catalyzers</i>	Fe_2O_3 , Iron(III) oxide or ferric oxide
	FeO n(OH), Iron(III) oxide-hydroxide
	nBF, n-butyl ferrocene
	LiF, Lithium fluoride
	CuCr_2O_4 , copper chromite

11.2.3 Solid Rocket Motor; Internal Ballistics and Ideal Performance

The fundamental relationships of SRM operation are still dominated by semi-empiricism. It is theoretically conceivable to solve the multi-phase problem of a reacting solid mixture coupled with the reacting gases produced at the gas–solid interface. In fact, most of the reactions that take place in the solid, a heterogeneous mixture comprising different-size particles of oxidizer salt and metal fuel inside a solid polymer, in the presence of catalyzers and additives, are unknown. Even if known, their interaction would have to be modeled at the scale of the finest particles, say 10 micrometers. As this approach is unfeasible, the fundamental Vieille’s law replaces actual unknown kinetics, as introduced in Chap. 8, thus

$$r = a(p_c)^n \quad (11.5)$$

where r is called the regression velocity, reported in cm/s or mm/s, and p_c is the chamber operating pressure. Note that Eq. 11.5 is a repetition of Eq. 8.40. The burning rate, or regression velocity, r , may vary between a few mm/s to more than 40 mm/s for certain applications. The factor a , known as the temperature coefficient, and the exponent n , sometimes called the pressure component or combustion index, must be experimentally found as a function of temperature and solid propellant composition; for instance by testing at fixed pressure in a test vessel a solid propellant cylinder ('strand') burning only in the manner of a cigarette. The mass rate of gas produced by solid propellant burning is given by

$$\dot{m} = A_b r \rho_p \quad (11.6)$$

where A_b is the area burning and ρ_p is the solid propellant average density prior to burning. Assuming a steady state, and hence neglecting any storage rate of hot gas in the combustion chamber, this mass rate must also be equal to the mass passed through the nozzle

$$\dot{m} = A_b r \rho_p = p_c (A_t / c^*) \quad (11.7)$$

where p_c is the chamber operating pressure, A_t is the throat area, and c^* is the characteristic velocity, introduced in Chap. 8 and is proportional to the absolute chamber temperature (known from the thermochemistry of the propellant); see Eq. 11.15. From this balance the steady state pressure inside the port of the SRM can be found as

$$p = [r_p c^* (A_b / A_t) a]^{1/(n-1)}. \quad (11.8)$$

Thus, the exponent n must be <1 for stable combustion; in practice it is between 0.3 and 0.5. Note also that (A_b / A_t) is an important dimensionless motor parameter with typical values much greater than 1 and is denoted $K \equiv (A_b / A_t)$.

To predict performance, the classic one-dimensional ideal rocket relationships are used. These assume isentropic, non-viscous expansion and a perfect or ideal gas. For an ideal gas, enthalpy can be expressed using the specific heat at constant pressure, c_p , and the absolute temperature; note that formally the specific heat at constant pressure is actually the partial derivative of the enthalpy with respect to the temperature at constant pressure. Assuming an adiabatic, no shaft-work process and the absence of shocks or friction such that the flow enthalpy change is zero, the total or stagnation enthalpy per unit mass, h_0 , is constant

$$h_0 = h + \frac{v^2}{2J} = \text{constant} \quad (11.9)$$

where J is the mechanical equivalent of heat, which is used only when thermal units, such as the British thermal unit

(Btu) or calorie, are mixed with mechanical units, such as the joule (J). In SI units the value of J is one and is neglected henceforth. The conservation of energy for isentropic flow between two sections shows the decrease in enthalpy as

$$\Delta h = \frac{1}{2} V^2 = c_p (T_c - T) = \left(\frac{\gamma}{\gamma - 1} \right) R T_c \left[1 - \left(\frac{p}{p_c} \right)^{(\gamma-1)/\gamma} \right] \quad (11.10)$$

where the subscript c refers to stagnation combustion (chamber) conditions, and γ is the specific heat capacity ratio, or adiabatic index, also denoted by κ (chemical engineers) or k (mechanical engineers); $\gamma = (c_p / c_v)$. The specific heat capacity ratio is a constant for perfect gases over a wide range of temperatures. The term in square brackets on the right-hand side of Eq. 11.10 is the thermodynamic efficiency, and R is the gas constant obtained either from the ratio between the universal gas constant, R' , and the average molecular weight, MW , of the exhaust gases, or from the difference between the specific heat at constant pressure and the specific heat at constant volume, i.e. $R = c_p - c_v$. The expansion area ratio for a nozzle with isentropic flow can be expressed as a function of the local Mach number, M , as

$$\frac{A}{A_t} = \frac{1}{M} \left[2 \left(1 + \frac{1}{2} (\gamma - 1) M^2 \right) (1 + \gamma) \right]^{(\gamma+1)/2(\gamma-1)} \quad (11.11)$$

where subscript t refers to the nozzle throat. These relationships hold at each p , T and V during an isentropic expansion due to varying cross-section area A .

The thrust, F , may be written as

$$F = p_c A_t C_F \quad (11.12)$$

where the thrust coefficient, C_F , introduced in Chap. 8, is a function of the specific heat capacity ratio, $\gamma = (c_p / c_v)$, and of the pressure ratio between chamber and nozzle exit pressure (p_c / p_e). That is

$$C_F = \Gamma \left[\frac{2\gamma}{\gamma - 1} \left[1 - \left(\frac{p_e}{p_c} \right)^{(\gamma-1)/\gamma} \right] \right] + \left[\frac{A_e}{A_t} \left(\frac{p_e}{p_c} - \frac{p_a}{p_c} \right) \right]. \quad (11.13)$$

Note that Eq. 11.13 is a repetition of Eq. 8.29, written in a slightly different form.

The ambient pressure, p_a , is practically zero in space. The thrust coefficient, C_F , is dimensionless and is typically found experimentally using Eq. 11.12 with measured values of chamber pressure, throat diameter, and throat, but it can also be found in textbooks as a function of the area ratio, pressure ratio and γ . Defining

$$\Gamma = \sqrt{\gamma} \left(\frac{2}{\gamma + 1} \right)^{(\gamma+1)/[2(\gamma-1)]} \quad (11.14)$$

the characteristic velocity, c^* , assumes the compact form

$$c^* = \frac{\sqrt{RT_c}}{\Gamma}. \quad (11.15)$$

The characteristic velocity, c^* , is a measure of the velocity (or of the I_{sp}) potentially available from expanding exhaust gas from the combustion temperature T_c to zero T and p . The mass flow rate, \dot{m} , passing through the nozzle becomes

$$\dot{m} = \frac{p_c A_t}{c^*} \quad (11.16)$$

Note that the term discharge coefficient, C_D , is occasionally used. It is simply the reciprocal of c^* . The thrust, F , may now also be written as

$$F = C_F \dot{m} c^*. \quad (11.17)$$

Finally, neglecting the contribution due to the pressure difference at the nozzle exit, that is, assuming the nozzle exit pressure is ‘adapted’ to the external ambient pressure, the specific impulse, in units of velocity, is

$$I_{sp} = \sqrt{\left(\frac{2\gamma}{\gamma-1} \right) \left(\frac{RT_c}{MW} \right) \left(1 - \left(\frac{p_e}{p_c} \right)^{(\gamma-1)/\gamma} \right)}. \quad (11.18)$$

These expressions can be used to estimate the performance of both SRM and liquid rocket engines, as in each case the flow is mostly gaseous. The major difference between the two is the molecular weight, MW , of the exhaust gas, which is higher for solid propellant than for most of the liquid propellant combinations.

Combustion of solid propellant is a feedback process similar to that which keeps a candle burning: the gas-phase flame pyrolyzes the solid fuel that releases the reactants that burn in the gas-phase. Once the solid propellant of a SRM ignites, the ammonium perchlorate (AP) and aluminum react exothermically and produce smaller hydrocarbon (HC) fragments that react with chlorine dioxide species in the gas-phase, raising temperature to 2,800–3,500 K. The hot gas feeds back heat to the surface that pyrolyzes and closes the energy cycle. The propellant surface is at a much lower temperature than the gas, about 800–1,100 K, while the gas flame may reach 3,500 K. High gas temperature coupled to the gas velocity (that increases in the port going downstream due to mass addition) may transfer too much heat to the surface and cause so-called erosion (too rapid and irregular surface regression). Erosion causes excessive pressure and thus a faster burning rate, r , and is one of the causes of combustion instability. Instability means rough combustion with pressure waves of large amplitude. In fact,

pressure spikes may reach twice the design pressure, with damaging or destructive effects. More generally, all chemical propulsion, not only solid propellant, is susceptible to instability. The reason is that combustion is a complex set of phenomena, including vaporization, pyrolysis, diffusion, convection, and chemical kinetics. Each of these has its own range of characteristic times, and when two or more overlap, coupling may occur, with one phenomenon in phase with, and reinforcing, the other and vice versa. Thus, combustion instability occurs also in liquid rocket engines and in hybrid rocket engines, see Sects. 11.3 and 11.4.

Although Vieille’s law is empirical and is an average derived from steady combustion measurements, for instance by strand burner tests, it is customary to apply it instantaneously and at each point of a burning solid propellant surface, especially in trying to predict dynamic or unstable combustion. This permits calculation of the mass flow rate, \dot{m} , and then the specific impulse, $I_{sp} = F/\dot{m}g$.

The shape of the burning surface A_b determines the $F = F(t)$ history and thus the trajectory and acceleration of the spacecraft. The initial shape changes with time, and to produce a thrust F that is progressive, regressive, neutral, or a combination (see Fig. 11.3), the burning area must be precisely predicted using internal ballistics, implemented in proprietary codes. End burning, as in a cigarette, gives a neutral F , while a cylindrical port yields a progressive F . To have a large A_b and a neutral curve the cross-section of the port must be tailored to remain almost constant in time. An example is the ‘dog bone’ in Fig. 11.3; at the burn-end the pressure decreases due to insufficient surface area burning, the burning rate r drops, and the solid propellant extinguishes, leaving unburnt propellant ‘slivers’ stuck to the case walls. A typical curve for the Ariane 5 booster is shown in Fig. 8.13.

An example SRM is given in Table 11.3, showing the key features of the MAGE family of SRM used on the first three series of the early Ariane launch vehicles.

The Zefiro family of SRMs developed for the Vega launcher by AVIO are a more recent example. Zefiro Z9 is the SRM powering Vega’s third stage, and is detailed in Table 11.4. The solid propellant binder is carboxy-terminated polybutadiene (CTPB). The case is filament-winding carbon-epoxy, with epoxy resin-based (EPDM) thermal protection. The nozzle is carbon-phenolic with an ablative carbon-carbon throat insert.

11.2.4 Solid Rocket Motor Manufacturing

Most of the technology of solid rocket motors deals with manufacturing. Unlike liquid rocket engines, SRMs using composite propellant are the result of many complex operations that need only to be summarized here. Starting with aluminum, ammonium perchlorate (AP) and a pre-polymer

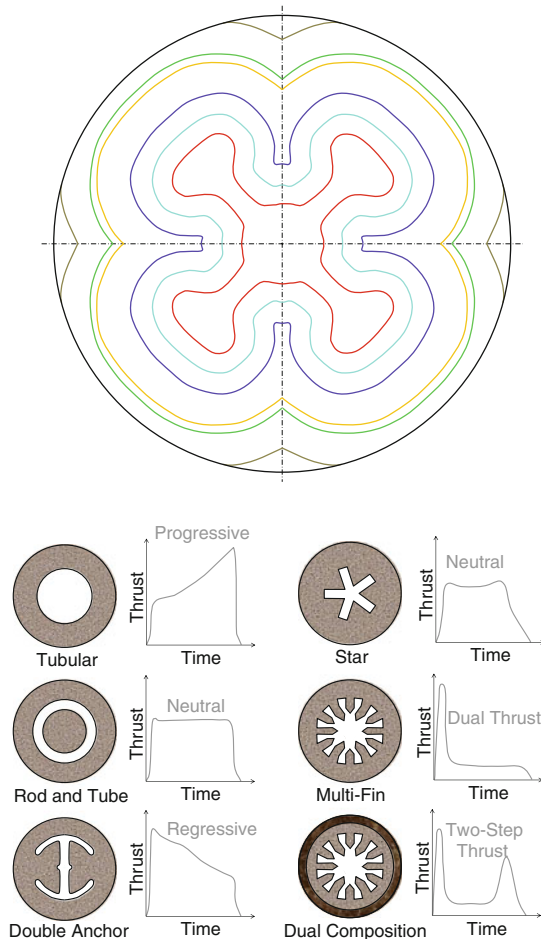


Fig. 11.3 Predicted time evolution of a solid propellant grain burning surface with a four 'dog bone' port cross-section (top), the initial shape of the port is in red; note slivers leftover at burn-end in tan color, and port cross-sections and their effect on the thrust history (bottom). Image Malcolm Macdonald

in liquid form, the first operation is to mix it with AP and aluminum powder. The rheological properties of the mixture are the key to a homogeneous propellant that will produce predictable and repeatable performance when burning. Once the case is fabricated, the thermal protection system (for instance, virgin rubber) is applied to the interior wall. The solid propellant mixture is then cast inside it and autoclave-cured at a specified temperature for hours or even days, for large grains. After curing, the grain is separated from the mold, and milled to ensure port area and shape match specifications. Testing samples, qualifications and many other operations follow that add to the cost of SRM.

11.3 Liquid Propellant Rockets

Liquid rocket engines (LRE) are propulsion systems that work by mixing and/or reacting one or more liquid propellants inside a combustion chamber, and ejecting the high

Table 11.3 Main features of the MAGE SRM family; the first figure in the propellant name is the aluminum percentage, the second is that of the binder

	MAGE 1	MAGE1S	MAGE2
Overall length (mm)	1,130	1,252	1,525
Outside diameter (mm)	766	766	766
Exit cone diameter (mm)	480	520	633
Propellant mass (kg)	272	410	490
Propellant	CTPB 16-12	CTPB 16-12	CTPB 16-12
Off-loading (kg)	+63 -22	-82	-90
Mean operational pressure, MEOP (MPa)	4.075	4.4	5.02
Maximum thrust (kN)	28.2	33.5	46.2
Burn time (s)	34	48.5	43
Expansion area ratio	45.5	54	65
Specific Impulse, I_{sp} , vacuum (s)	284.7	291	294
Nozzle material	Carbon-carbon and carbon-phenolic	Carbon-carbon and carbon-phenolic	Carbon-carbon
Case material	Filament winding Kevlar [®] 49	Filament winding Kevlar [®] 49	Filament winding Kevlar [®] 49
Operating temperature (°C)	-10 to +40	-10 to +40	-10 to +40

pressure exhaust through a thermodynamic nozzle. The momentum acquired by the exhaust gases produces the thrust, F .

Monopropellant liquid rocket engines are typically used to control the attitude of a spacecraft, to maneuver (for instance, to adjust and change orbit) and to deorbit. A monopropellant reacts by decomposing exothermically, for instance by passing through a catalytic bed. Examples of monopropellants are hydrazine (N_2H_4), hydrogen peroxide (H_2O_2) or mono-methyl-hydrazine (MMH). The fuel is forced out of the tank using a stored inert gas, typically nitrogen or helium. If the gas is stored in a separate tank, the system is 'regulated' and the gas tank will be at 3,000–8,000 psi, while the fuel tank may be at 200–300 psi; a regulated monopropellant system schematic is shown in Fig. 11.4. If the gas is stored in the same tank as the fuel, the system is 'blowdown' and the tank pressure will range

Table 11.4 Zefiro Z9 key parameters

Parameter	Value
Length	3.7 m
Diameter	1.9 m
Solid propellant mass	10.5 metric tons
Case mass	388 kg
Nozzle expansion ratio	60.8
Nozzle throat diameter	0.16 m
Average thrust, F	276 kN
Specific impulse, I_{sp} , vacuum	294 s
Burn time	106 s
Maximum burn pressure	75 bar
Area ratio	56

from 300 to 400 psi at beginning-of-life (BoL) to approximately 100 psi at the end-of-life (EoL); a blowdown monopropellant system schematic is shown in Fig. 11.5.

Bipropellant liquid rocket engines comprise two propellant tanks (storing fuel and oxidizer), lines feeding the propellants in an assigned mass mixture ratio to the rocket combustion chamber (the ‘chamber’ for short), and a system to increase the pressures of the propellants for combustion because high pressures increase the thermodynamic efficiency of the rocket cycle. Common bipropellants are liquid oxygen/liquid hydrogen (LOX/LH₂), LOX/hydrocarbons (HC), where HC may be kerosene or RP-1 (Rocket Propellant-1 or Refined Petroleum-1), liquid methane (LCH₄), liquid propane (LC₃H₈), and hydrazine (N₂H₄) or mono- or dimethyl-hydrazine (MMH or UDMH, respectively) and dinitrogen tetroxide (N₂O₄); typically referred to simply as nitrogen tetroxide (NTO). Further, NTO is often used with the addition of a small percentage of nitric oxide, which inhibits stress-corrosion cracking of titanium alloys; in this form, propellant-grade NTO is typically referred to as mixed oxides of nitrogen (MON) and is typically the more common form of NTO for spacecraft propulsion. By example, a reaction control system may use NTO containing 3 % weight solutions nitric oxide (3 wt % NO), termed MON3. In the past nitric acid (HNO₃) and other toxic combinations were also used [2]. Recently, so-called ‘green’ combinations have been proposed and tested, such as LOX/ethanol and H₂O₂/kerosene. They are called green because they are neither toxic nor produce toxic products; however, LOX and H₂O₂ still need great care in handling.

Maintaining chamber pressure can be done simply by pressurizing the propellant tanks, or by using turbopumps for the fuel and for the oxidizer. The power to drive the turbopumps is obtained by pre-burning a small fraction of the propellants in a gas generator and expanding the hot products in a turbine. In most gas generators the mixture is rich, in order to avoid chemical attack by the oxidizer. A notable

exception are the staged combustion, RD-170 and RD-180 (РД-170/180, Ракетный Двигатель-170/180, Rocket Engine-170/180), engines developed in Russia which burn oxygen-rich in their gas generator; the RD-180 is shown in Fig. 11.6 during a test firing. Typically, the turbine shaft drives both turbopumps, but a gearbox has previously been used to drive two pumps at two different speeds. This can be required for cryogenic engines, as the mass flow rate of oxidizer and its density are invariably much larger than that of the fuel. This difference can be such that the two different flow rates cannot be sufficiently realized by simply sizing the centrifugal pump stages appropriately.

Cryogenic engines tend to use two separate turbopumps, with each mixture ratio controlled by the respective turbine flow. The feeding pressure controls the mass that burns in the combustion chamber at each instant, and thus the chamber pressure, p_c . As the space vacuum pressure is close to zero, liquid rocket engines for use in space need a p_c of only a few tens of bars to achieve reasonable thermodynamic efficiency. Figure 11.7 shows a schematic of a regulated bipropellant rocket engine.

Turbopumps build enough head to circulate one of the two propellants in order to cool the engine. Generally, the coolant is the fuel rather than the oxidizer, in order to prevent chemical attack; circulating propellant keeps the engine walls at approximately 600–800 K, while combustion within the thrust chamber produces gases at temperature of order 2,500–3,500 K, which no material would otherwise be able to withstand at typical operating pressures (50–100 bars).

For low thrust engines, the tank pressure may be insufficient to achieve good circulation, because forcing propellant inside the jacket of a liquid rocket engine can result in a significant pressure drop. When the chamber and nozzle are cooled only by radiation it is often manufactured from a single piece of ceramic or refractory metal(s), for instance niobium or rhenium; see for instance Fig. 11.8.

The Aerojet engine in Fig. 11.8 has a radiatively-cooled chamber. The area ratio was 129, with a specific impulse of 348 s, and thrust, $F = 44.5$ kN. This engine was designed for the ascent from the lunar surface of the future Lunar Surface Access Module. Figure 11.8 shows that actual expansion in the test cell was less than that possible *in vacuo*.

Fuel and oxidizers are injected inside the combustion chamber through the so-called injector plate by means of separate passages. These may be coaxial ducts, where the central pipe carries liquid oxidizer, and the annular duct carries fuel. Non-cryogenic liquid propellants, for instance, hydrazine and NTO, are injected through angled orifices: the high-speed jets impinge on each other, splash, form droplets and burn upon coming into contact. Droplets vaporize much faster than jets, and combustion can initiate. Hydrazine or MMH or UDMH and NTO are ‘hypergolic’

Fig. 11.4 Schematic of a regulated monopropellant system

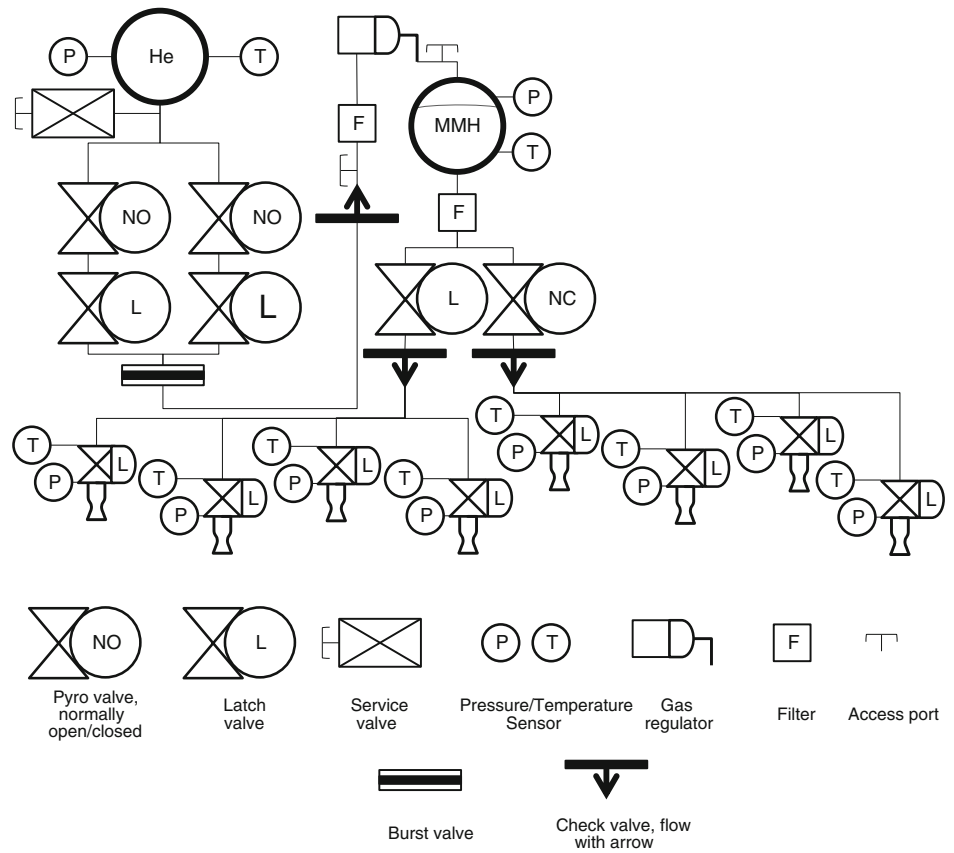
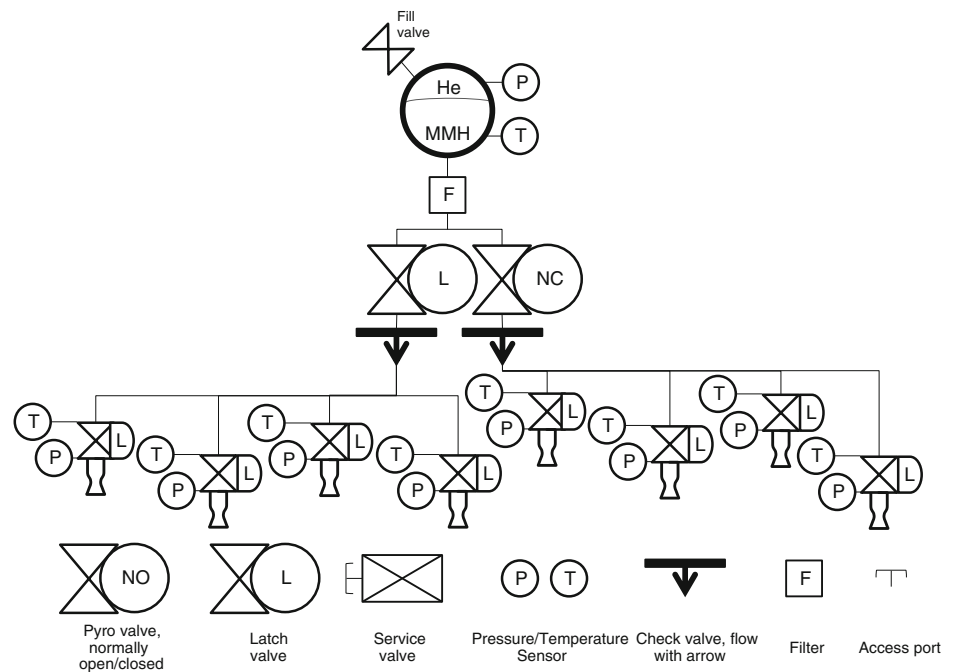


Fig. 11.5 Schematic of a blowdown monopropellant system



combinations, and as such they do not need ignition systems. Most common propellant combinations such as LOX and LH₂, or LOX/hydrocarbons (HC), are not hypergolic. Depending on the size (thrust) of the liquid rocket engine,

they need ignition devices that range from simple spark plugs to miniature rocket engines and torches. An overly long ignition delay after startup (after starting injection) means accumulation of propellants in the chamber and



Fig. 11.6 Test firing of the Atlas III propulsion system configured with the Russian-designed RD-180 engine on November 4, 1998 at the Marshall Space Flight Center (MSFC) Advanced Engine Test Facility. The RD-180 is powered by kerosene and liquid oxygen. *Image NASA*

‘hard starts’, i.e. formation of pressure peaks that may damage or even destroy the engine.

To operate in space, liquid rocket engines must be fed propellants irrespective of the direction of acceleration or spacecraft attitude. The mechanical design of each tank requires the propellant collector to be always immersed in liquid, even in weightlessness, when acceleration is not present, and even when the propellant is ‘sloshing’ due to a spacecraft maneuver, for instance, driven by auxiliary (attitude control) thrusters. If pressure-fed, the pressurizing strategy is based either on a polymeric or metallic membrane surface that is capable of staying flexible at the propellant temperature and of withstanding possible chemical attack, or on propellant acquisition by surface tension, for example, with LOX/NTO. Polymeric membranes are mostly used with hydrazine for short- or medium-duration operation. The pressurizing gas (say, helium) compresses the membrane on one side, which in turn pressurizes the liquid propellant on the other side. Bipropellants (NTO/MMH) use generally surface tension tanks.

The same feeding problem occurs in space operations with liquid rocket engine turbopumps. In addition, due to the minimal tank pressurization (2–5 bars) each pump must be capable of ‘sucking’ the propellant flow demanded by the engine during startup without cavitating (making bubbles). As the oxidizer and fuel flows must optimize engine performance, the turbopumps maintain the mixture ratio, MR , constant from startup to final burnout. This is especially critical with low density, supercritical LH_2 ; the solution here consists of an ‘impeller’, shaped like a multi-blade outboard propeller with very coarse or helicoidal

pitch that precedes the more conventional centrifugal pump stage. It is however of note that in the case of the hydrogen upper stages of the Saturn V (at least), the mixture ratio was ‘PU shifted’ (propellant utilization shifted) part way through the burn, once the actual performance of the engine had been determined, to ensure that the fuel and oxidizer were used up at the same time.

A few tripropellant liquid rocket engines have been proposed and even ground-tested. Examples include the $LC_3H_8/LH_2/LOX$ system proposed by R. Beichel while at Aerojet, where LC_3H_8 is the fuel during liftoff, and LH_2 at altitude; the NPO Energomash RD-701 which uses kerosene/ LGH_2/LOX ; and even $F_2/H_2 + Be$. The design complication, toxicity and logistics of fluorine-containing propellants have prevented their utilization, although the specific impulse would be higher than bipropellant systems; a tripropellant system may have a specific impulse of >540 s, whereas a LOX/LH_2 bipropellant system *in vacuo* may attain a peak specific impulse of around 465 s.

11.3.1 Engine Cycles

Liquid rocket engines are thermodynamic machines and follow thermodynamic cycles dominated by pressure. Propellants are compressed, burn at constant pressure, and expand in the nozzle to produce the desired power, $P = FV$, where V is the velocity of the vehicle. This is a simple Brayton-type cycle, but if turbopumps and a gas generator are present, the cycle becomes more complicated. Chamber pressure and the expansion ratio (see Sect. 11.2) determine the cycle efficiency; the maximum cycle pressure must be higher than the chamber operating pressure, p_c , to enable the propellants to be injected at high speed and mix properly, and thus require a certain ‘head’. A further constraint is imposed by combustion dynamics (instability), which dictates a minimum required pressure drop Δp through the injector system in order to decouple unavoidable pressure oscillations in the chamber from the feeding system. In fact, the flow regime in the propellants feeding ducts is typically subsonic, and in the absence of throttling between the chamber and the tanks, drives what is called ‘pogo instability’. This Δp may be of order 10–50 bars, depending on the chamber pressure and injection system.

The thermodynamic efficiency of space propulsion is a function of the area ratio (A_e/A_e), where subscript e is the nozzle exit, or pressure ratio, (p_c/p_e), and can be raised by simply lengthening the conical (divergent) part of the nozzle, as the ambient pressure is zero, and (up to a point) it is more convenient to increase A_e and thus lower p_e than it is to raise p_c . The area ratio may even reach 400 with composite or ceramic nozzle extensions or skirts that can move axially and form a longer, larger area ratio nozzle to increase expansion

Fig. 11.7 Schematic of a regulated bipropellant system

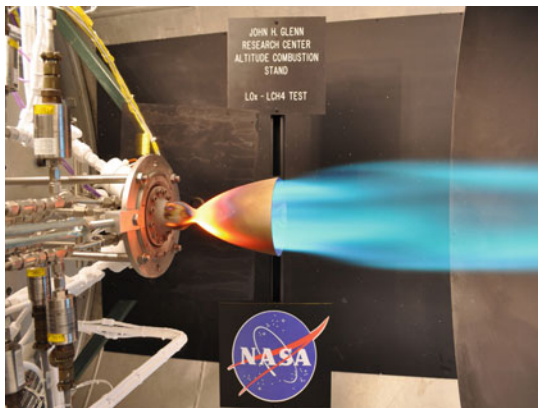
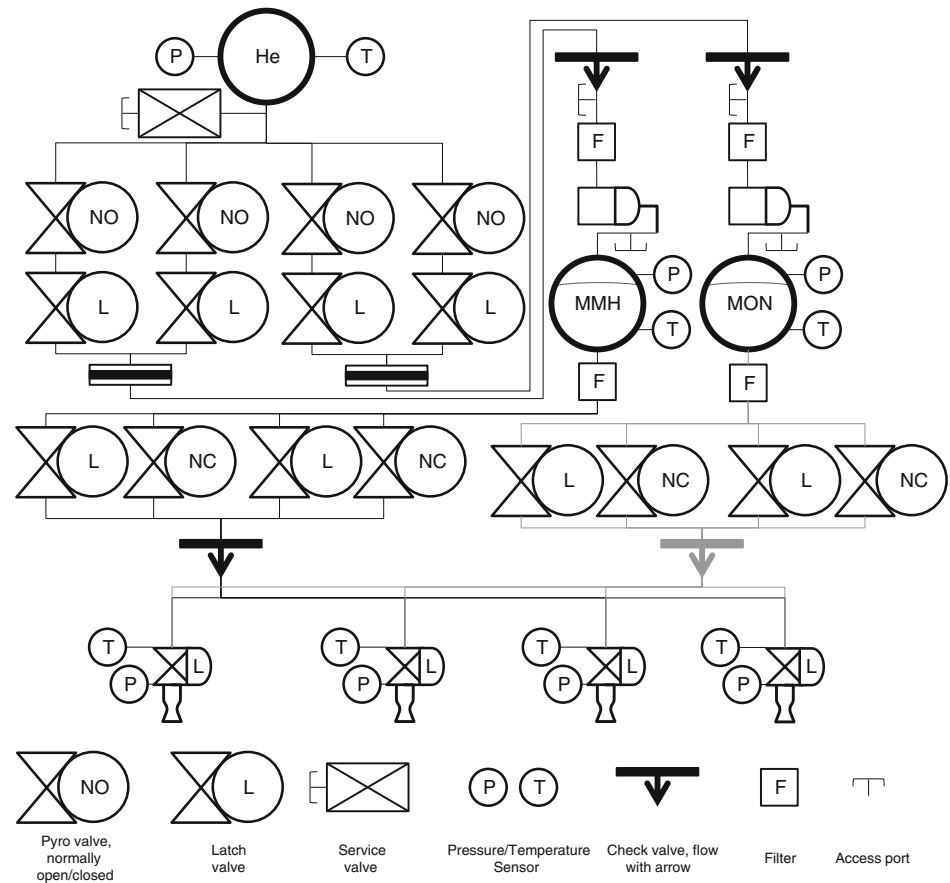


Fig. 11.8 Prototype of Aerojet LOX/LCH₄ (liquid oxygen/liquid methane) space engine during ground tests at the Glenn Research Center. This liquid rocket engine is radiatively cooled, with the heat flux peak near the nozzle throat as shown by the radiance intensity

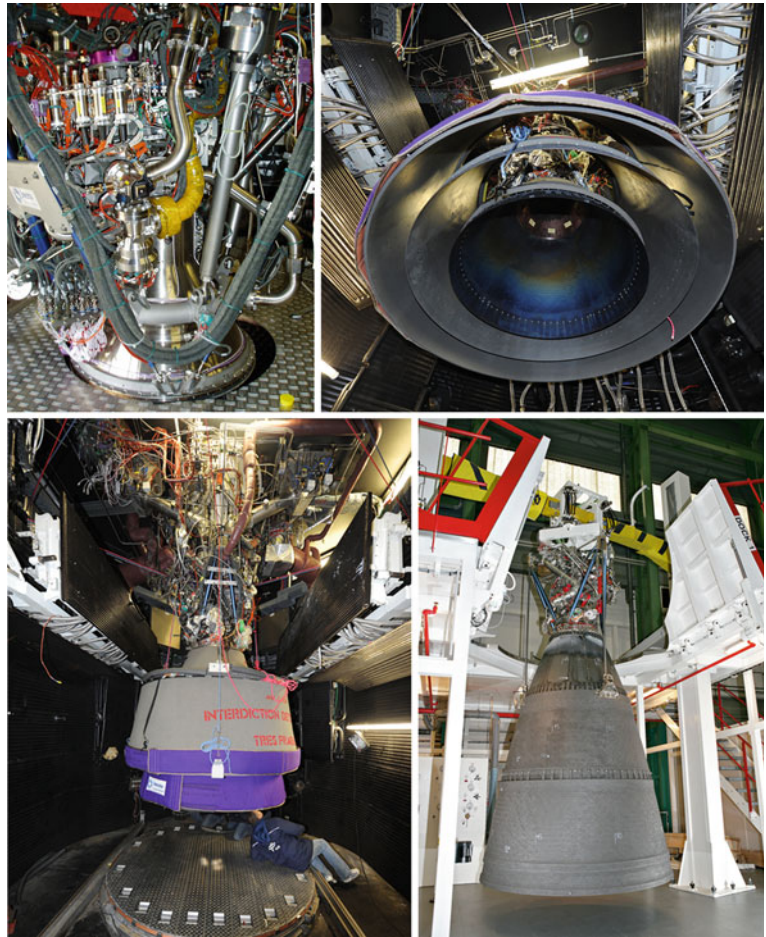
at altitude. This number should be compared to, say, 40–60 for the liquid rocket engine of a launcher. In space applications, p_c may be limited to <100 bars, but some space engines work at higher chamber pressure. The means of producing pressure dictates the cycle.

The cycles of bipropellant liquid rocket engines may be open or closed. Open cycles pre-burn a small amount of

propellants (usually with a rich MR in order to limit exhaust gas temperature) in a gas generator that drives a single or separate turbines which in turn, drive the propellant turbopumps. The turbine exhaust is ejected to space or injected into the nozzle divergent. Thus the fuel remaining in the rich turbine exhaust is wasted. In open cycles the turbopumps need only to produce a head that is the sum of $(p_c + \Delta p)$, the pressure loss that is necessary to decouple the feed system from the chamber and prevent self-sustained pressure oscillations. This Δp can be a fifth or sixth of the chamber pressure.

A special case of open cycle for in-space LOX/LH₂ propulsion is the so-called expander cycle; it creates pressure by using LH₂ to regeneratively cool the nozzle and chamber. The heat extracted vaporizes LH₂, forming gaseous H₂ to drive the turbopumps. An example is the Vinci engine, see Fig. 11.9, developed as the first European reignitable cryogenic upper stage engine for the Ariane 5 launcher. The LOX turbopump speed is 18,000 rpm, while the LH₂ turbopump speed is 90,000 rpm, delivering turbine powers of 350 kW (LOX) and 2,800 kW (LH₂); the propellant flow rates are 33.7 kg/s (LOX) and 5.8 kg/s (LH₂). Vinci produces 180 kN of thrust in a vacuum by burning LOX/LH₂ at $MR = 5.8$, with area ratio 240 (ceramic skirt fully extended). The chamber operating pressure, p_c , is

Fig. 11.9 Vinci reignitable cryogenic upper stage liquid engine. The expansion nozzle, is made of a carbon-fiber ceramic composite material and weighs 130 kg. It is shown both in its stowed configuration with a protective shell surround the expansion nozzle and in its deployed configuration; 2.37–4.2 m height, nozzle exit diameter 2.2 m. *Image ESA*



about 61 bars and the specific impulse 465 s. As the power is simply due to the expansion of LH_2 to GH_2 , the power generated is much less than in a gas generator cycle, and the chamber pressure is limited to a few tens of bars, but the engine is much simpler, less expensive, and more reliable.

Closed cycle liquid rocket engines, also called staged-combustion cycles, have the equivalent of the gas generator, now called a pre-burner, feeding combustion gas to the turbine (or turbines). In this case, however, the exhaust, still fuel rich, is not ejected overboard but injected into the combustion chamber, so that all of the fuel is eventually burned. The downside is that in order to inject the turbine(s) exhaust into the chamber the turbopumps must create a much higher head than in an open cycle. Because the turbine must produce much more power to drive the turbopumps, the pre-burner supplying the thermal power becomes almost as important as the main combustion chamber. Compared to open cycles, closed (staged) cycles are much more complex. Their higher specific impulse is at the price of higher thermo-mechanical stresses and weight, so they are less common for in-space propulsion, where reliability is a key requirement.

Chamber pressure is a compromise between thermodynamic efficiency/performance, and cost. Both the performance

and weight of a liquid rocket engine increase with its chamber pressure, and the crossing point depends on the propellant combination, the technology available, and most importantly, on the mission.

11.3.2 Engine Cooling

Engine cooling is critical to liquid rocket engine operation. The local heat flux inside a chamber increases from the injector plate to the throat as the surface area shrinks, and then decreases in the diverging nozzle. The throat is the most thermally loaded part of a liquid rocket engine, as seen in Fig. 11.8, with fluxes inversely proportional to the radius of the throat and of order MW/m^2 . Liquid rocket engine cooling may be active, with fuel being circulated inside the walls of the engine, or passive, by radiation. The engine walls may be actual pipes, assembled and brazed, as seen in Fig. 11.10, to form the combustion chamber jacket. This is an efficient cooling strategy, but is also expensive because each pipe each must be tapered near the throat. This feature maintains the propellant velocity near the hotter parts of the engine, and at the same times increases the mass flux, a

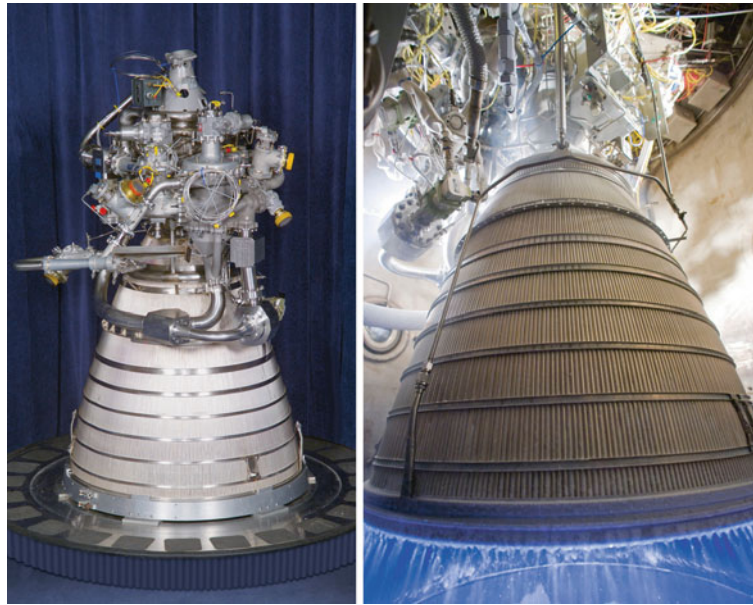


Fig. 11.10 The Pratt & Whitney-Rocketdyne Common Extensible Cryogenic Engine (CECE), based on the design of the heritage RL-10 engine, a deep-throttling 15,000 pound (~ 67 kN) thrust technology development engine fueled by a mixture of 90 K liquid oxygen and 20 K liquid hydrogen. The engine components are super-cooled to

similarly low temperatures. As the CECE burns its frigid fuels, gas composed of hot steam is produced and propelled out the nozzle to create thrust. The steam is cooled by the cold engine nozzle, condenses and eventually freezes at the nozzle exit to form icicles. *Image Pratt & Whitney-Rocketdyne*

desirable feature since the cooling requirement increases. A cooling jacket is more economical. In this case, the jacket is formed by an internal and an external surface, with brazed spacers between the two surfaces shaping coolant channels as required by the heat transfer. European liquid rocket engine technology typically uses cooling channels machined from a solid copper block and closed by a thick electrolytic nickel layer. In addition to circulating propellant, it is common practice to inject fuel through slots or orifices on the injector plate close to the engine sidewall in order to keep the wall wetted by a film of cold fuel and thereby insulate it from the combustion gases.

Radiative cooling is more practical for low thrust engines. Attitude control thrusters, monopropellant liquid rocket engine, and generally engines with a thrust of less than 10^4 N are not actively cooled; they may be manufactured using refractory metals such as niobium or rhenium, or more recently also ceramics such as silicon carbide (SiC), silicon nitride, trisilicon tetranitride (Si_3N_4), and carbon or SiC composites. The power radiated is a net energy loss, but that can be partially made up by the greater expansion ratio available. In any case, all the complications of manufacturing and operating active cooling systems are entirely avoided. Expander cycle engines have limited circulation capability, so only the chamber proper may be actively cooled; the diverging part of the nozzle and the nozzle extensions, if any, may be made of ceramic or composite, and cooled by radiation.

Another form of passive cooling is based on ablative materials. Single-use liquid rocket engines may have the inner walls of their chamber coated with poorly conducting, silica-reinforced epoxy or phenolic resins. The coating may be of the order of a few millimeters thick, decreasing over time due to the endothermic reactions between the ablative and the combustion gases. While the ablating surface reacts and is consumed, the thermal wave penetrates only a fraction of its thickness, maintaining the underlying structure at a reasonably low temperature. Ablative coatings may work more than once, but because their performance is not 100 % predictable they pose a risk if the engine is required to be restartable.

11.3.3 Liquid Rocket Engine Operation

Even at the ‘high’ specific impulse of 450–460 s, the working time of a space liquid rocket engine is limited to several minutes due to propellant consumption. For example, an RL-10 engine at 110 kN thrust and specific impulse 465 s burns its propellants at the rate $(F/I_{sp}) = 24$ kg/s.

Missions with liquid rocket engine propulsion often require engine restartability in order, for instance, to circularize an orbit or to inject the spacecraft into a Hohmann trajectory, or to make a planetary landing. Multiple restart (re-ignition) is sometimes critical, for instance in interplanetary missions. Hypergolic propellants react as soon as they come into contact, and can be restarted any number of times.

Liquid rocket engines with non-hypergolic propellants need reliable ignition systems built into their design from the very beginning. Thrust control is a second critical issue for some missions. For landing on bodies such as the Moon or Mars, the engine must be throttleable. While most liquid rocket engines may be capable of varying their thrust between, say, 95 and 102 % of nominal, there are several engines whose thrust may be reduced by a factor >1 . One such example is the Common Extensible Cryogenic Engine (CECE) engine built by Pratt & Whitney-Rocketdyne for lunar landings; see Fig. 11.10. Alternatively, the thrust may be controlled by pulsing periodically the engine, or (much more expensively) by subdividing thrust among many individual chambers, and igniting as many as are needed to obtain the thrust desired.

11.3.4 Operational Problems

Liquid rocket engines suffer from instability problems due to interactions among the physical mechanisms that control heat release. Cryogenic engines with gaseous hydrogen and LOX injection are supposed to be free of combustion instabilities, but are not always so. On the other hand, instability is a serious problem with the liquid/liquid injection scheme (typically, with hypergolic mixtures). Instability manifests itself by chamber pressure oscillations around the nominal value in time and space, there being longitudinal, tangential and radial oscillation modes. Some of the modes and their harmonics are acoustic, but there may be finite-amplitude modes, especially following startup and hard starts in general. All modes tend to cause problems, for instance by sweeping the wall boundary layer and thus sharply increasing the transfer of heat to the chamber wall. The presence of a liquid phase due to atomization of both fuel and oxidizer adds to acoustic modes generated by the effect of pressure waves on combustion kinetics. In-space liquid rocket engines tend, as a class, to be smaller than the first- or even third-stage engines of a launch vehicle, so harmonics and unstable modes tend to have higher frequencies, and sometimes that is beneficial, because it makes them more difficult to excite. Nevertheless, instability is an unpleasant fact with liquid rocket engines, and its abatement and control consumes most of the ground testing time.

Nozzle non-equilibrium effects are similar to those in launchers, but the lower pressure in the nozzle expansion slows further radical recombination and causes performance losses. In a nozzle, the hot combustion gas expands and cools, converting chemical and thermal energy into kinetic energy. Moving downstream, the gas accelerates, and there is less and less time for this conversion. Eventually, at a certain station of the nozzle the rate of pressure and temperature drop becomes faster than the rate of conversion, the flow composition and temperature stay unchanged, and the

gas flow is said to be chemically ‘frozen’: from that station on, the fraction of internal energy that is not converted into kinetic energy is lost.

A second source of losses occurs in very small engines (micro-thrusters for certain small satellites), because viscous effects become important below certain sizes (say, at Reynolds number, Re , below 100). The Reynolds number decreases along a nozzle because as the gas velocity increases, its density decreases faster. For such small engines, the radial velocity profile is no longer flat, it becomes parabolic, and the cross-section average velocity is less than that predicted by the classic one-dimensional rocket relationships. Realistic predictions of performance require simulations of the engine flow-field in detail.

11.3.5 Propellants and Performance

The one-dimensional rocket equations in Sect. 11.2 show that liquid rocket engine performance depends on the thermodynamic efficiency (pressure ratio), which in turn depends on the geometry of the engine (area ratio) and on the ratio between the combustion temperature, T_c , and the average molecular weight, MW , of the combustion products. The pressure ratio depends on the propellants only in terms of the specific heat capacity ratio, $\gamma = (c_p/c_v)$, but their chemical kinetics determine the ‘frozen’ losses in the nozzle. Other key considerations in choosing propellant combinations are their combined, or bulk, density (density weighted with the propellants mass fractions), the mixture ratio because it impacts on the design and size of the turbopumps, their possible toxicity, corrosiveness, logistics and handling problems, and cost. Even after much trial and error since the 1940s, current bipropellant combinations are relatively few, the most successful being LOX/LH₂, LOX/kerosene, NTO/N₂H₄ and NTO/MMH. Recent emphasis on ‘green’ propellants has raised interest in fuels such as ethane and ethanol. Dozens of combinations have been tried and used for some period in the past, and the discussion below may clarify why so few remain in use.

11.3.5.1 Liquid Oxygen/Liquid Hydrogen (LOX/LH₂)

As predicted by Tsiolkovsky, in space the LOX/LH₂ combination is the best performing, with a theoretical specific impulse of 465 s for an area ratio of 200. Both propellants are cryogenic. LH₂ requires to be kept at about 20 K and it took a long time for the US and the former Soviet Union to master this technology. The other countries that have succeeded include France, Japan, China, and India. LH₂ is about a factor of 20 more expensive than kerosene. The stoichiometric reaction $H_2 + \frac{1}{2} O_2 \rightarrow H_2O$, produces about 59 kcal/mol and has a $MR = 8$. The actual MR to achieve the maximum

specific impulse in space is about 6. Because this mixture is about 30 % rich, the adiabatic combustion temperature is not the highest obtainable, but the average molecular weight of the exhaust, which is about 10 due to the unburnt hydrogen, is much lower than for the stoichiometric mixture (about 18). LH_2 is supercritical when pumped and is therefore an excellent engine coolant. The heat that it extracts gasifies the liquid and raises its temperature to 250–300 K. The gaseous H_2 is then injected by means of coaxial injectors through the larger cross-section annulus coaxial to the central LOX ‘post’. The lower density of gaseous hydrogen (GH_2) results in injection velocities of the order of 100 m/s, while the denser LOX enters the chamber through its central post at only a few meters per second. The energy yield and cooling are outstanding qualities for a liquid rocket engine, especially for the first stage of a launcher. However, LH_2 cannot be stored in space for long periods, as it evaporates. Active cryocoolers have been investigated, but the ratio between the heat extracted and the energy used to extract it is currently still of the order of 1/30 to 1/100. Hydrogen may be kept liquid only by active cooling or by steady evaporation, at a rate that depends on tank insulation and on exposure to sunlight. At a warm-side temperature of 350 K, good multilayer insulation transmits 0.28 W/m^2 , and because LH_2 tanks are bulky (the density of LH_2 at 20.4 K is only about 0.07 kg/m^3) boil-off is significant. Advances in insulation strategy have made passive storage competitive in space for times up to 60 days, and for a 5–10 t cryo upper stage (1–2 t of LH_2) in-space active cryocooling of LH_2 has been demonstrated with a power consumption of less than 1 kW (10 W at 20 K). Poor space storability and low bulk density are the main disadvantages of LH_2 as a space propulsion fuel. With LOX, this combination is ‘green’, producing only water. Ground handling is relatively benign, but the wide flammability limits of hydrogen and oxygen mixtures always pose a risk, prevented in practice by appropriate safety measures and operating procedures.

11.3.5.2 Liquid Oxygen/Hydrocarbon (LOX/HC)

The combination LOX/kerosene is better suited to boosters and launchers because it is denser, having a bulk density of about 1.4 relative to water. LOX is partially space storable, as it vaporizes below its critical pressure (50 bars); kerosene is fully storable. This combination delivers a specific impulse of approximately 357 s with a nozzle area ratio of 40. The combustion products are mainly CO_2 and water; both are considered ‘green’, but because RP-1 contains cycloalkanes and aromatics, it is slightly toxic, although less so than gasoline. Very similar specific impulses are obtained with other hydrocarbon fuels, for instance liquid methane (LCH_4) and liquid ethanol (LC_2H_6), which are much cheaper than hydrogen and have better handling characteristics and logistics.

LOX/kerosene engines need an ignition system, hence engine restarting is problematic. Furthermore, any fuel that remains in the ducts after turning the engine off pyrolyzes quickly to form carbon deposits (‘coking’). To some extent, this happens with all hydrocarbons. Coking depends on combustion temperature and pressure, impurities in the fuel (no commercial hydrocarbon is pure), and catalytic reactions with metallic surfaces. In the worst cases, valves may be clogged after only a few minutes of operation. Kerosene used in the US for rocket propulsion is a special blend (military specification MIL-P-25576C) called RP-1 (Rocket Propellant-1 or Refined Petroleum-1) and is more expensive than commercial kerosene; for comparison, at the time of the Apollo program the cost was 1.45 US\$/gallon versus 0.27 US\$/gallon for gasoline (1 gallon is approximately 4.546 l). The RP-1 blend was formulated to reduce coking while maintaining a specific density of 0.81–0.82 in order to reduce tank volume. The average carbon to hydrogen ratio of RP-1 is about 1.953.

For optimum specific impulse the *MR* is 2.77 with kerosene, 3.45 with LCH_4 and 3.10 with C_2H_6 (ethane). Significant experience exists in Russia and Ukraine with LOX/kerosene combinations. LCH_4 was used in the former Soviet Union, and is currently undergoing testing both in the US and in Europe as an alternative to both kerosene and hydrogen. Its properties are intermediate between the two: LCH_4 has far better cooling capability than kerosene, and far less than hydrogen; it is much denser than LH_2 , but less so than kerosene, and it has less tendency to coke. Injection of LOX/RP-1 in the Rocketdyne F-1 engine of the Saturn V for Apollo was through like-on-like and like-unlike impinging jets; in Russian designs coaxial ducts are more common.

11.3.5.3 Hypergolic Combinations

Hypergolic combinations in current use are based on hydrazine or its mono- and unsymmetrical dimethyl compounds (MMH and UDMH) as fuel, and nitrogen tetroxide (NTO) as oxidizer. These propellants can be stored for years, but NTO must then be supplemented with nitric oxide (NO). The combustion products depend on the hydrazine compound. With straight hydrazine (N_2H_4) they are mainly N_2 and H_2O , but N_xH_y species may be present. With MMH and UDMH, CO_2 and H_2O are also main products. Both propellants are dense liquids at room temperature, and both are hazardous and toxic. Their specific impulse performances are very similar, about 340 s with an area ratio of 40. The optimum *MR* is about 2.4 for MMH, 2.15 for UDMH and 1.4 for neat hydrazine. Because they are all hypergolic, pipe joints and valves must be designed with high tolerances to prevent leaks and catastrophic single-point failures. For the same reason, multiple restarts pose no problem because ignition and extinction are performed by simply opening and closing the propellants valves. These

Table 11.5 Some properties of commonly used space propellants

Propellant	Critical temperature (K)	Critical pressure (bars)	Boiling temperature (K)	Liquid density (g/cm ³)	Heat of vaporization (kcal/kg)
LH ₂	69	12	20	0.071	107
LO ₂	115	50.1	90	1.14	50.9
LCH ₄	192	45.8	112	0.423	122
RP-1	675	21.4	450–547	0.82	74
Hydrazine (N ₂ H ₄)	653	145	387	1.004	300
MMH	585	75	361	0.866	183
UDMH	523	49.3	337	0.971	126
NTO (N ₂ O ₄)	431	91.4	185	1.55	96

combinations are preferred for spacecraft attitude control and orbital maneuvering. Fuel and oxidizer are injected from pressurized tanks through separate orifices on the injection plate. The liquid jets are angled to make them impinge at a distance from the plate sufficient to protect the plate from overheating. Hydrazine, MMH and UDMH are all relatively stable with temperature, and can be circulated to regeneratively cool the engine walls without decomposing. Table 11.5 summarizes some properties of the most common propellants discussed.

11.3.6 Monopropellants

The two most commonly used monopropellants are hydrazine (N₂H₄) and hydrogen peroxide (H₂O₂).

11.3.6.1 Hydrogen Peroxide

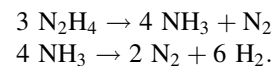
To optimize performance, H₂O₂ must contain as little water as possible. In fact, the heat released by its decomposition (H₂O₂ → H₂O + O) is capable of vaporizing completely the peroxide at a concentration exceeding approximately 70%. Above 85%, H₂O₂ is called high-test peroxide (HTP). Anhydrous peroxide is difficult and very expensive to produce; the highest available concentration is about 98%, and its density is 1.44 g/cm³.

HTP tends to decompose in tanks and release heat: in fact, it reacts with most metal or impurity traces. These catalyze the reaction H₂O₂ → H₂O + O, although less so at the highest concentration. Very pure aluminum tanks are best, and special additives (e.g. tin salts) have been found that slow decomposition even further, to the point that HTP can be used for orbital maneuvering and attitude control; compatibility with catalyst must be always checked. The same reaction is desirable inside the catalytic reactor, which may even consist of a simple silver screen, but that will melt readily because the temperature of the decomposition reaction is about 1,543 K. Platinum and other proprietary

catalysts (e.g. based on iron oxides) are also used. The specific impulse obtainable is of order 189–195 s, and decomposition starts as soon as the valve opens and lets the peroxide wet the catalytic bed. HTP is toxic, but the real hazard is its high reactivity, producing burns when in contact with any type of organic material.

11.3.6.2 Hydrazine

Hydrazine (N₂H₄) is historically the first and by far most reliable monopropellant. Its shelf life inside certain types of stainless steel, or aluminum alloy tanks, can be measured in years, and experience exists over decades of use in attitude control systems, station-keeping for geostationary satellites and orbital maneuvering. A hydrazine motor consists of an injector, a catalyst bed, and a chamber ending in a conventional nozzle. Over the catalyst bed, the exothermic decomposition of N₂H₄ takes place in two steps, the first exothermic and the second endothermic i.e.



The second step (ammonia decomposition) is slow compared to the first. The longer the residence time over the catalyst, the more ammonia (NH₃) is converted to nitrogen and hydrogen. This fact can be exploited to control the trade-off between specific impulse performance and the combustion temperature, T_c . At 6.8 bars pressure, and an area ratio 50, a short catalyst bed produces ammonia and a maximum specific impulse of 260 s, but the temperature is 1,700 K. Dropping the catalyst temperature to 1,273 K prolongs catalyst life considerably but reduces the specific impulse to 245 s. The catalyst lasts longer at lower temperatures. Increasing the bed/residence time so that all ammonia is converted can reduce the combustion temperature to about 870 K and the catalyst life is accordingly even longer, but the specific impulse is only 210 s. The engine designer must therefore choose appropriately the bed length and injection velocity.

As with H_2O_2 , the thrust can be controlled simply by the opening and closing of a valve. Fast acting valves can deliver thrust pulses very precisely and reliably; that is, with minimum dribbling. For attitude control, hydrazine thrusters may consist of groups of three or even four independently fueled thrusters (to increase reliability), oriented in the three main spacecraft or satellite axes. A single fuel or oxidizer tank feeds all of the thrusters, but their valving is independent.

The current emphasis on ‘green’ propellants has produced a possible non-toxic competitor to hydrazine, called LMP-103S by its manufacturer, ECAPS of Sweden. It is composed of ammonium dinitramide, water, methanol and ammonia. It decomposes at higher temperatures than hydrazine into water, N_2 , H_2 and CO/CO_2 , with a specific impulse of about 265 s. It was successfully tested on the Swedish Mango and Tango satellites launched in 2010. The catalyst life is for the moment shorter than with hydrazine.

11.4 Hybrid Rockets

Hybrid rocket engines (HRE) are a class of propulsion systems that were introduced in the 1930s; work on the GIRD-9 (Soviet) LOX/jellified gasoline 60 lbf (267 N) thrust motor built by Mikhail Tikhonravov (1900–1974) and Sergei Pavlovich Korolev (1907–1966) began in 1932 and its first flight in August, 1933. The hybrid rocket engine concept is to burn a solid fuel with a liquid oxidizer. Most of the fuels tested are polymers, such as polyethylene (PE), hydroxy terminated polybutadiene (HTPB), which are also used in solid propellants, or methyl-metacrylates (commercially: PlexiglasTM). Typical oxidizers are LOX and H_2O_2 .

The reason for developing hybrid rocket engines is their historically perceived greater simplicity than both solid and liquid systems. Compared to a SRM, a hybrid motor hosts a simpler and cheaper fuel grain (not a propellant grain), and its thrust can be controlled by regulating the oxidizer flow rate; and compared to liquid rocket engines, a HRM has a single liquid propellant tank and a single turbopump (or pressurization system). The working pressure of a hybrid rocket engine is a few tens of bars, and is determined by the injection pressure. With a simple polymer fuel, the combustion products consist of CO_2 and H_2O . SpaceShipOne was powered by a hybrid rocket using HTPB and nitrous oxide (N_2O), and that has focused attention to this oxidizer. With N_2O there will be also N_2 and perhaps traces of amines among the combustion products. N_2O is a gas that can be liquefied relatively easily for storage and is non-toxic, but decomposes exothermically once sufficiently heated.

Theoretically, the specific impulse of hybrid rocket engines is intermediate between that of liquid and solid systems, in the 300 to 330 s range, depending on the oxidizer. Thus, in principle, hybrid rocket engines for in-space

propulsion and their inherent ‘restartability’ would facilitate multiple burns at lower cost. Figure 11.11 shows a notional hybrid rocket engine. This comprises the liquid oxidizer tank, the line feeding oxidizer to the rocket chamber via a conventional injector, and the fuel grain inside the chamber. Even for large-scale applications, turbopumps would pose major problems, so the tank is simply pressurized. This limits the operating chamber pressure to a few tens of bars.

Similar to SRMs, the solid fuel is shaped as a grain with a central port whose cross-section is shaped to control the total surface/volume ratio for optimum combustion. The polymeric fuel containing C_xH_y groups such as CH_2 or CH pyrolyzes due to the radiation from the flame and due to convection. The products of pyrolysis react with the oxidizer, which is LOX, N_2O , or H_2O_2 . The reaction occurs where the mixture fuel/oxidizer is stoichiometric, thus at a certain distance from the surface. The surface regresses much like in a SRM, except the mechanism is driven exclusively by the heat feedback to the fuel surface, and is slower. The regression rate of commonly used polymers is much smaller than that of solid rocket motor propellant, a few mm/s at most.

Hybrid rocket engines are still in a protracted development phase, and should be rated perhaps more in terms of their safety and cost than their performance. They suffer from several basic shortcomings. Ideally, combustion should take place near or at the stoichiometric mixture layer. However, because of the time needed to vaporize it, the oxidizer may mix with the pyrolyzed fuel while still in the form of a spray of droplets. This is typically the case near the entrance of the port, and reactivity is low there. In addition, fuels like HTPB or polyethylene pyrolyze slowly. Thus, the burning rate peaks toward the grain end, and much fuel (10–15 %) may remain unburned. This is also due to the difficulty of ensuring that the liquid not only vaporizes, but also reaches the entire fuel surface, and in the right amount, while it atomizes and is convected downstream. This limits the effective thrust. Increasing the motor surface area by multiple ports increases at the same time its volume. In space that does not matter, but the motor has to be lifted into orbit, and a bulkier payload implies a bigger aerodynamic fairing, more drag and more structural weight. Also, low-frequency instability is common during hybrid rocket engine operation.

11.4.1 Hybrid Rocket Engine Evolution

Much of the research in hybrid rocket engines has gone into ways of increasing their thrust by increasing the contact between liquid and solid fuel. An obvious way is to increase the port area, shaping the grain cross-section as a wagon wheel, for instance. This creates multiple independent ports. However, any multi-cavity configuration of the fuel grain tends to weaken it mechanically, and the regression rate

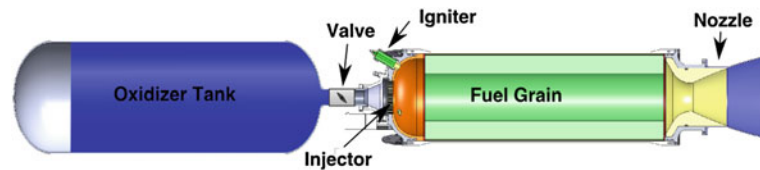


Fig. 11.11 Notional scheme of a hybrid rocket engine. The liquid oxidizer is pressure-fed to the injector. *Image Creative Commons*

tends to be non-uniform from port to port due to differences in the oxidizer flow rate delivered to each port.

Combustion in hybrid rocket engines occurs via a classic diffusion flame where the fuel is produced by the flame heat feedback, as in an ordinary candle. It is the heat feedback from this flame that pyrolyzes the solid fuel by transferring heat to the solid surface by a combination of radiation and also convection. Thus, the burning rate is controlled by the rate of inter-diffusion of fuel and oxidizer that must first mix in order to react.

To increase the burning rate, and thus reduce the grain bulk and increase the thrust for a given volume, necessitates the use of one of the following strategies

- Increase the surface area of the fuel in contact with the oxidizer
- Increase heat transfer between hot gas and surface
- Increase reactivity between gas and oxidizer.

The first approach is similar to, or the same as, that used for SRMs. The grain must be milled to have as much of its cross-section consisting of empty space (ports) where the oxidizer may enter and react with the fuel as it pyrolyzes. There is an obvious compromise between the surface area that can be achieved and the structural grain integrity during combustion and spacecraft acceleration. Furthermore, as the grain is consumed, the walls in between ports get thinner, the burning inside the different ports is uneven, and eventually slivers of fuel detach and are carried downstream, toward the nozzle throat. Once removed, they rarely burn completely. In addition, a multi-port cross-section increases the burning surface at the expense of the bulk density (density/unit bulk volume). For this reason hybrid rocket engines tend to be more voluminous than SRMs, and launching them is more expensive due to the extra drag.

The second approach is based on increasing the oxidizer turbulence in the port(s) so that the turbulent layer over the fuel surface, where combustion occurs, becomes thinner, and thus the heat transfer is faster. The mixture ratio to produce the optimum specific impulse is fixed once the chamber pressure and nozzle have been fixed, as is the oxidizer mass flow rate, which is equal to the product of the velocity, cross-section, and density. The Reynolds number that determines the intensity of turbulence scales with the product of these three factors. It may be increased only by

reducing the port area, but as the fuel surface decreases so will the flow rate of pyrolyzed fuel. Thus, this approach has limitations if done conventionally. Recent work has produced more innovative ideas. An example is single-port grains, where the oxidizer swirls and more rapidly and more completely consumes the fuel, since the gaseous boundary layer swept by the swirling flow becomes much thinner. This concept is being developed by ORBITEC in the US, and is called ‘vortex hybrid’. The disadvantage is that multi-port grains would need multiple swirling oxidizer injectors, one to each port, thereby increasing both complexity and cost. In practice, this approach is currently suitable only for small thrust hybrid rocket engines, but it may eventually find a niche in space propulsion. A second innovative approach, developed at University of Hokkaido, Japan, divides the fuel grain into segments, with empty spaces in between. Each segment has two or three ports, but these are staggered with respect to the next segment so that the hot combustion products are forced to recirculate before passing through the next set of ports. As in the swirling strategy, it is this recirculation that increases the rate of heat transfer to the fuel and its pyrolysis.

The third approach is newer. It consists of replacing the polymers with much more easily vaporized waxes (e.g. solid paraffin). The endothermic heat of pyrolysis of common waxes is a fraction of that for polymers, so more fuel is available to the oxidizer per unit time, and the diffusion-controlled combustion is faster. This has been shown to have a positive effect. For instance, solid waxes leave less unburnt fuel (slivers). However, compared to polymers, waxes have much weaker mechanical properties, especially in hot environments. In space that may become a problem if exposure to the heat of solar flux is prolonged.

A further concept is to use fuels that are ‘doped’ with oxidizing agents or metal particles. Metal particles increase the flame temperature and radiate intensely inside the flame. Metal nano-particles work best, but they are expensive, and mixing them into the fuel poses manufacturing problems. Nano-particles of metals that burn to oxides with a high adiabatic flame temperature, are boron, aluminum and zirconium and their hydrides. Hydrides release hydrogen when burning, thus contributing to energetic kinetics and total heat release. With this strategy, the supposedly ‘cheap’ fuel actually becomes similar to a solid propellant.

Although the first two strategies may increase combustion completeness, they also tend to increase the overall pressure drop, with a negative effect on thrust and specific impulse. The first strategy probably requires abandoning conventional designs. The turbulent transport of oxidizer to the surface can be obtained with unconventional designs that increase the recirculation zones and thus the residence time of the fuel and oxidizer mixture. This can be seen as transforming hybrid rocket engines from a propulsion system similar to a SRM, to something closer to a gas generator.

11.4.2 Hybrid Rocket Engine Burning

Empirical formulations of the regression rate of the solid fuel are used in determining the performance and design of hybrid rocket engine. The regression rate, r , concept is used as in SRMs, but it is no longer a characteristic of the fuel because it depends on the details of the heat transfer from the boundary layer flame. The regression rate is correlated to the oxidizer ‘mass averaged velocity’, G_{ox} , defined as the oxidizer mass flow rate divided by the total port area, giving the regression rate in units of distance by time as

$$r = a(G_{ox})^n \quad (11.19)$$

where a and n are empirical parameters. For instance, with the combination HTPB and LOX, $a = 0.104$ and $n = 0.681$ (here r is in inches per second, and G_{ox} in pound-mass per second per square inch, $\text{lbm}/(\text{s in}^2)$). The value of the regression rate depends on the fuel and on the operational conditions; experiments show regression rates of 1–4 mm/s, but these numbers are merely indicative because heat transfer and pyrolysis depend on the specifics of the motor, not just on the propellants. With solid paraffin fuel $a = 0.488$, and although $n = 0.62$, the regression rate is a factor of three faster. Pressure dependence (expected by the combustion mode, based on a diffusion flame) is implicit in G_{ox} . However, other empirical relationships for the regression rate include an explicit chamber pressure dependence, with $r = a(G_{ox})^n (p_c)^m$.

As with other chemical engines that operate at a multitude of characteristic times which may couple, hybrid rocket engines are subject to instabilities. Unlike SRM and liquid rocket motors, hybrid motors have not been observed to self-destruct due to excessive pressure growth. Pogo, or chugging instability due to coupling between chamber pressure and tank feed system, can be cured by a suitable pressure drop across the liquid injector, as in liquid rocket engines. Acoustic instabilities are more difficult to cure as in the case of liquid rocket engines, and indicate that the acoustics of the ports and of the motor case need to be modified.

11.4.3 Hybrid Rocket Engine Performance

The American Rocket Company (AMROC) tested hybrid rocket engines, with thrusts from 44 to 334 kN, including a very large H-1800 motor in the early 1990s to demonstrate practically the capability of hybrid rocket engines to replace solid boosters; it developed a thrust of 1.1 MN during a single run, but exhibited instability and combustion was incomplete. Later, in 1999, a consortium of aerospace companies, working under NASA’s Hybrid Propulsion Development Program, tested a motor with the same thrust, again for application to boosters. The nominal chamber pressure was 61 bars, and the test lasted 80 s. Combustion roughness was observed. These were the largest hybrid motors ever tested; they demonstrated both the advantages and problems of this type of propulsion.

The peak specific impulse of LOX with HTPB fuel *in vacuo* is obtained theoretically for $MR = 2$ and is about 330 s; with H_2O_2 the optimum MR is 6, and the specific impulse is about 300 s. These are ideal calculated numbers. In 2002, Lockheed-Martin flight-tested its Hybrid Sounding Rocket (HYSR) sounding rocket, powered by a hybrid rocket engine with a thrust of 270 kN.

The most celebrated application of a hybrid rocket engine was in the suborbital SpaceShipOne vehicle built by Scaled Composites and SpaceDev, which won the X-Prize in 2004. This hybrid rocket engine used liquid N_2O and HTPB, and produced 88 kN of thrust. Similar to H_2O_2 , the N_2O decomposed exothermically, and there was a risk of the flame flashing back from the injector to the tank. A flashback was responsible for the explosion of the hybrid rocket at Scaled Composites in 2007 that killed three people.

Hybrid motors are subject to low-frequency (1–100 Hz) instability. However, the pressure does not grow catastrophically, as it can in liquid and solid systems. ‘Chugging’ is fairly common at low oxidizer flow rates, when turning the thrust down. With LOX, the main mechanism for instability is recognized to be the time lag of vaporizing the LOX droplets, just as in a liquid rocket engines.

Although still not mature, hybrid rocket engine technology shows promise where *safety*, not performance, is the key requirement.

11.5 Electric Propulsion Fundamentals

Electric propulsion is a technology based on accelerating matter by means of electric forces. To do that, the matter must first be ionized. The forces that can accelerate it may be electrostatic (Coulomb) or electrodynamic (Lorentz). The first needs only an electric field, i.e. a voltage difference, and the force between two point-like charges q_1 and q_2 at distance r is

$$\mathbf{F} = Kq_1q_2 \frac{\mathbf{r}}{r^3} \quad (11.20)$$

where the constant $K = 1/(4\pi\epsilon_0)$. The vacuum permittivity ϵ_0 is 8.854×10^{-12} Farads per meter (F/m). The Lorentz force acts in the simultaneous presence of an electric and a magnetic field, and for a single charge, e , it is simply

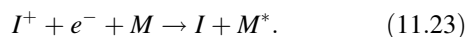
$$\mathbf{F} = e\mathbf{U} \times \mathbf{B} \quad (11.21)$$

where \mathbf{U} is the velocity of the point-like charge e , and \mathbf{B} is the magnetic induction. For a fluid mixture composed by $i = 1, 2, \dots, N$ ionized species, with charge density $q_i\rho_i$, subject to both an electric field, \mathbf{E} , and a magnetic field, \mathbf{B} , the total electric (body) force acting on the i th component of the mixture is

$$\mathbf{F} = \rho_i q_i (\mathbf{E} + \mathbf{U}_i \times \mathbf{B}) \quad (11.22)$$

This force must be added to the Navier-Stokes momentum equations. Note that the Lorentz force depends on velocity, so is not Galilean invariant, and this motivated the use of the Lorentz transformation in Special Relativity. A rigorous simulation of the effects of electric forces on a charged fluid with charge density $q_i\rho_i$ would require adding *all* of Maxwell equations and *all* of the boundary conditions. The complexity of electric thrusters has resulted instead in shapes and practices that by trial and error over half a century have yielded good or promising results for thrust and specific impulse. Thus, this technology is still in a development phase, and there are no simple relationships to predict thrust and specific impulse with good accuracy. Most often, power given as input is empirically correlated to thrust by direct measurements.

To accelerate ionized gas, the gas density must be sufficiently low that collisions do not recombine ions and electrons. Thus, the residence time in the region inside an electric thruster using, for instance, the Coulomb force, must be shorter than the average time between collisions. Recombination between an ion I^+ and an electron e^- occurs through a three-way collision with any third body M present



M is any third body capable of absorbing the recombination energy, and its state M^* indicates that after the three-body collision, M leaves in an excited state. The rate of recombination is thus proportional to the concentrations of all three collision partners simultaneously, and therefore scales with pressure cubed: the higher the pressure, the larger is the probability that ions and electrons will recombine. This limits the maximum pressure of electric thrusters based on the Coulomb force.

Similarly, if the thruster uses the Lorentz force, the trajectories of ions and electrons, for constant and uniform \mathbf{E} and \mathbf{B} , will be spirals around the magnetic field lines of constant value. If the spiraling period (gyration time) is shorter than the collision time then the Lorentz force will have time to act and will accelerate the ions and electrons; otherwise, the ionized gas is said to be ‘collisional’ and the kinetic energy imparted by the Lorentz force will be ‘dissipated’ through inelastic collisions and become heat. The ratio between the gyration time and the collision time in the gas is the so-called Hall parameter. In ionized gas the percentage of ions may be seemingly ‘low’, say 10 %, but the high velocity given to ions will be distributed among all molecules present (ions, electrons and neutrals). A gas does not need to be 100 % ionized in order to be efficiently guided and accelerated by electromagnetic forces.

These fundamental considerations explain why electric thrusters are capable of accelerating ions to speeds that are impossible using chemical reactions, and that whilst they are capable of high specific impulses, their thrust comes nowhere near that of chemical rockets: because the pressure must be very low, so is the momentum of the ionized gas being accelerated, and therefore the thrust.

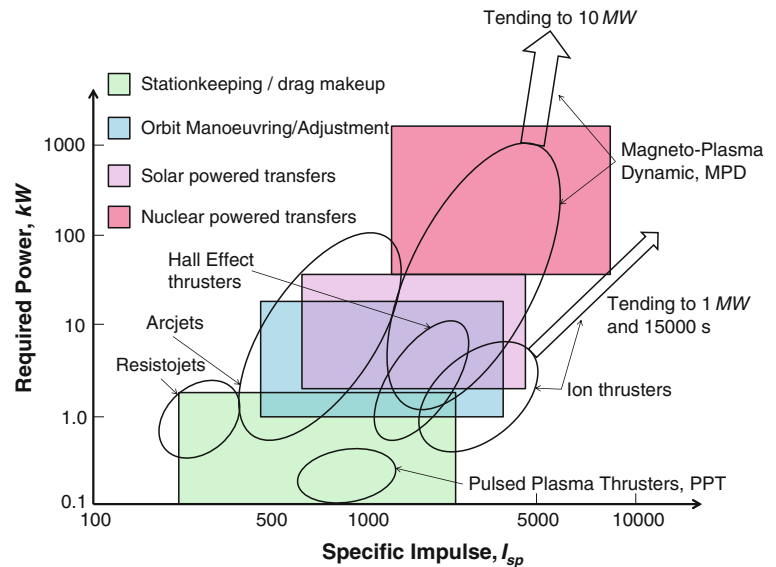
Electric thrusters (ET) where the accelerating force is Coulomb are called ion thrusters, and if the acceleration is driven by differences in voltages applied to grids at the exit of the chamber, the thruster is often called a gridded ion engine (GIE). The jet of ions is also called an ion beam. The ion beam is *de facto* a streamtube carrying a flow of ions, and thus is equivalent to a conductor carrying a current. Current is measured in ampere (coulombs per second), or more practically, in *mA* with gridded ion engines. Ion thrusters operate at thousands of volts, but their current density (the ion flow rate) is small. As an example, an ion current of 5 A is considered ‘large’.

The body of an ion thruster is light, consisting of a simple cylinder, or a conical chamber, in which the ions are produced at very low pressure. Their volume is large compared to the thrust they can produce. Although relatively lightweight, they are cumbersome propulsion systems.

Thrusters driven by the Lorentz force, or where a magnetic field holds some function, are of many types, and may be grouped together under the name of magnetoplasmadynamic (MPD) thrusters. MPD thrusters may have a self-generated magnetic field, that is, a magnetic field created by the ion current, I , itself, or an externally applied magnetic field; for instance, using permanent or electromagnets.

Applied-field MPD thrusters generally need high current (thousands of A) and low voltage (10 or hundreds of V). Ohmic losses, which scale with I^2 , are significant. The body of an applied-field MPD thruster is heavier than that of an ion thruster because conventional electromagnets require copper

Fig. 11.12 Approximate map of power and specific impulse available with different electric thrusters, adapted from [3]. Note the log-log scale



coils. This may change over time as superconductor materials replace copper, enabling the use of much smaller coils. Self-field MPD thrusters may be non-stationary (pulsed), and if so, they generally produce a low thrust. The thrust/unit volume may be ten times higher than that of a GIE.

An important effect to be avoided in electric thrusters is the charging of the engine itself during operation. For instance, in a GIE positive ions are accelerated and ejected, thus the thruster itself charges negatively with time as electrons accumulate in the engine. If electrons are not ejected as well, the voltage of the thruster will rise until arcing takes place between parts of the engine that are at different voltages. To avoid this, the accelerated ion jet must be neutralized: electrons must be conveyed and ejected *outside* the thruster towards the ion beam in order to make it electrically neutral.

Common to all electric thrusters is a mechanism to ionize the propellant. Gases such as O_2 , N_2 , H_2 , Ar, He, Xe, Ne, have ionization energies of their atoms 13.6, 14.5, 13.6, 15.75, 24.6, 12.1, 21.6, respectively. The ionization potential of an alkali metal is lower, e.g. Li has 5.39 eV, Cs has 3.89 eV, but to be practical these low ionization metals must be stored as liquids, and then vaporized to serve as propellants. Note that ionization is not the only factor in choosing a propellant, but the specifics of the ionization mechanism may determine the particular type of electric thruster. How the propellant is ionized characterizes different thrusters.

In arcjets (see Sect. 11.7) propellant is ionized by an arc discharge. In a GIE, the ionization may be achieved by a radio-frequency (RF) voltage, by microwaves or by electronic bombardment of the gas by electrons from a hollow cathode that spews electrons to the chassis (at a different potential from the cathode). MPD may use hollow cathodes, or RF heating.

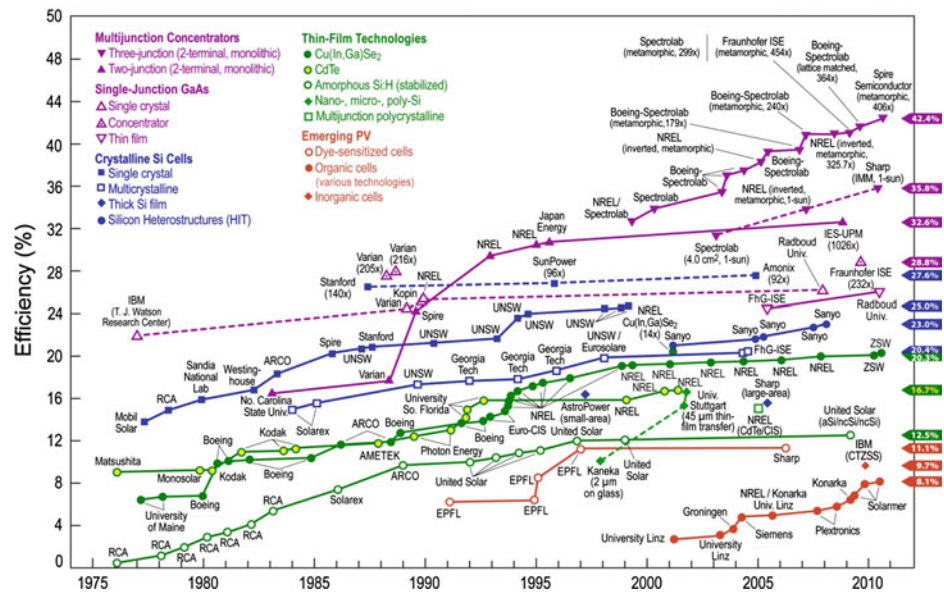
A power versus specific impulse map of electric thrusters is shown in Fig. 11.12, adapted from [3]. Note that the upper boundary of MPD thrusters could tend towards 10 MW, and that ion thrusters could tend towards 15,000 s and have been run in the lab at 1 MW.

11.5.1 Propellants

Choice of propellant depends on the type of electric thruster. In GIE, where pressure must be kept low, to increase thrust the momentum of the propellant should be maximized. Heavy molecular weight propellants are thus preferred. Note that this is exactly the opposite for thermochemical rockets, where thrust can be made as high as desired, and specific impulse is limited by temperature. Arcjets excepted, in all types of electric thruster it is an *external force* that accelerates ions, not a thermodynamic expansion, therefore higher molecular weight is convenient. The molecular weight of the propellant should be the highest that is compatible with operational and logistic requirements. These dictate molecules that are gaseous (or liquid) at room temperature, are relatively easy to liquefy and store in tanks, are safe to handle, and hopefully not too costly. For instance, mature GIE technology has xenon as the propellant of choice. Xenon is not cheap (about 4,000 US\$/m³), but its atomic number, Z , is 54 and its density at standard temperature and pressure is about 5.8 kg/m³. It is also inert and safe to handle. Its density when liquid (~ 3.1 g/cm³) means small tanks. Note that at supercritical pressure its density is lower. Historically, the first propellants used for GIEs were mercury and caesium.

In arcjets the pressure need not to be as low as in a GIE or MPD thruster. Its thrust is produced by thermodynamic expansion in a nozzle, as in a liquid rocket engine, and low

Fig. 11.13 Evolution of solar cell best research-cell efficiency with time. Chart includes commercial and research cells. Image NREL



molecular weight propellants like H_2 have been used successfully. In MPD, acceleration is achieved by an external (Lorentz) force, but H_2 has been used as a propellant because it can be ionized and accelerated to absorb much larger power than other gases. The discovery that Li has no second ionization losses and (in association with Ba) reduces cathode erosion, has produced a significant improvement in the performance of MPD thrusters. Applied-field MPD have also been tested with argon in Europe. The 200 kW VASIMR thruster that will run on either deuterium (D, or 2H), is due to be tested on the International Space Station (ISS) in the second half of the 2010 decade, although using argon for safety.

A common problem to all electric thrusters (but not to VASIMR) is electrode erosion and, more generally, material degradation when bombarded by ions and electrons. Other issues are engine bulk; power conditioning (e.g. GIE need high voltage, and current is typically low; MPD need just the opposite, as do arcjets); electric power supply, relying on solar panels that degrade with time due to solar and cosmic background radiation; efficiency, lower than 50 % in certain thrusters; and heat waste disposal.

The electric thrusters that can be bought off the shelf are GIE and Hall thrusters. Most high power MPD are promising but still in the developmental stage.

11.5.2 Power Supply

Electric thrusters typically use solar power. In low-Earth orbit, where the solar constant is $1,366 \text{ W/m}^2$, the initial conversion efficiency is of the order of 24–28 %. The remainder is thermalized, and heats the panels. Heating and the effect of solar proton flux degrades conversion

efficiency. Future solar panels may have a higher efficiency (see Fig. 11.13) but their technology is still maturing.

Large voltages are possible by linking panels in series (e.g. for ion thrusters). Parallel connections in order to produce the large currents needed by MPD thrusters, require much more wiring, which poses a weight issue that is independent of the panel weight.

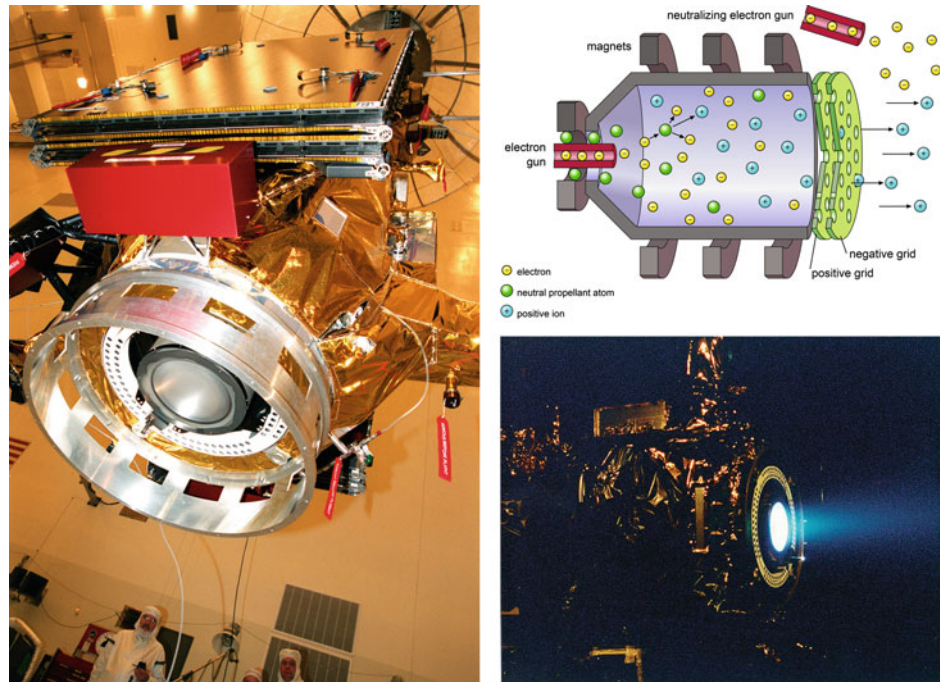
Depending on the type of cell, the power to mass ratio may vary between 40 W/kg for rigid panels and 140 W/kg for blankets. Per unit area, the power varies between 100 and 300 W/m^2 . A new technology uses solar concentrators over each cell, and may produce up to 290 W/m^2 , but also raises solar array mass. Further discussion of solar cells can be found in Chap. 10.

It is apparent that for interplanetary exploration, and in particular, for Mars missions, solar power may be an issue rather than a solution. This will depend on the mission duration, because crewed missions need fast interplanetary transit in order to reduce the radiation dose to the crew, and also because the solar flux available for power generation decreases with the square of the distance from the Sun. Thus, nuclear reactors will probably replace solar power for deep space missions to the external planets.

11.6 Electrostatic Propulsion

Electric thrusters using Coulomb forces to accelerate positively charged ions (ion thrusters) are in principle very simple; they were envisaged by Konstantin Tsiolkovsky, experimented with by Robert Goddard, and first tested by Ernst Stuhlinger (1913–2008) at Peenemunde during WWII. Harold R. Kaufmann (born 1926) at NASA-Lewis (now Glenn) was the first to use high-energy electrons to ionize

Fig. 11.14 NASA NSTAR Engine on the Deep Space-1 (left), and under testing (bottom left). A Schematic of an electrostatic ion thruster is also shown at top right; the ionization device is a hollow cathode, emitting and bombarding the propellant with electrons. The grid system consists of only two grids. Another hollow cathode emits the electron beam to neutralize the thruster and the ion beam at the same time. *Image NASA and Creative Commons*



the propellant gas (mercury) in a practical way. The two first thrusters in space operation, SERT-I and -II were successfully launched by NASA in 1964 and 1970. Their ideal application was found in north/south station-keeping of geostationary satellites, where trajectory changes are minimal and the thrust required is very low. This notwithstanding, the space industry is lukewarm to novelty, and only in the 1990s did they become accepted as a satellite propulsion systems. More than 100 Xenon Ion Propulsion Systems (XIPS), originally manufactured by Hughes, then by Boeing, and now by L3-ETI, are in operation. In the higher power class, NASA developed the NSTAR, which served as the main propulsion thruster of the DeepSpace-1 (DS1) and Dawn missions. Its thrust is about 92 mN but can deliver a total ΔV of several km/s.

A schematic of an ion thruster is given in Fig. 11.14. This shows the conical chassis, a system of two grids at different voltages used to accelerate ions, and a neutralizer. The presence of the grids has given in the UK the name gridded ion engine (GIE) to this type of electric thruster. The propellant (Xe is the most common) is ionized and a voltage difference is applied between the two grids. The chassis is at a different voltage from the first grid (the ‘screen’ grid), in order to accelerate electrons emitted by the hollow cathode. It is the electron collisions with the propellant that are responsible for its ionization.

The ions that are formed are accelerated by the difference in voltage (~ 1 kV) between the screen grid and the second (‘accel’) grid. The acceleration process of a partially ionized gas in an electric field is quite different from that of a single ion because positive charges repel each other, and

the thruster volume is filled with slow moving ions with a high charge density, thus substantially altering what would seem to be a straightforward task. In fact, the acceleration of ions takes place between the grids, not between the chassis and the grid(s). The purpose of the screen grid is simply to extract slow ions from the positive plasma sheath upstream, so that they may be accelerated by the voltage difference with the ‘accel’ grid. Guiding ions by means of voltages is similar to guiding light with lenses, and this fact has coined the term ‘ion optics’. More efficient schemes using three or four grids have been successfully demonstrated. Figure 11.15 shows the voltage scheme of the μ -10 Japanese microwave-ionization thrusters installed on the Hayabusa probe that sampled the Itokawa asteroid. This ion thruster used three-grid optics, with the chassis held at the same voltage as the screen grid.

11.6.1 Performance

The ideal zero-dimensional energy balance [4] predicts the ion velocity U (the specific impulse of the thruster) in the acceleration region. If the mass of the ion of charge q is m_q , this balance states $m_q(U^2/2) = q\Delta V$, where the right hand side is the work done by the Coulomb force on the charge q subject to the voltage difference $\Delta V = V - V_0$. Thus, $U = \sqrt{2(q/m_q)\Delta V}$. The flux of ions times their charge, is $j = nUq$, in A/m², where n , the number of ions per unit area, is the charge density in one-dimension, j is also the ion beam current density. To find the thrust F it is necessary to

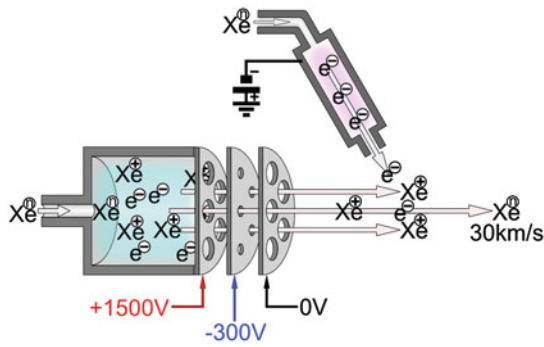


Fig. 11.15 Schematic of the voltages applied to the three-grid μ -10 ion thrusters of the Hayabusa probe that sampled the Itokawa asteroid and returned to Earth in 2006

calculate $V = V(x)$ under the effect of the space charge density n . This may be done by writing the differential one-dimension Poisson equation for V

$$\frac{d^2V}{dx^2} = \frac{nq}{\epsilon_0} \quad (11.24)$$

where $\epsilon_0 = 8.854 \times 10^{-12}$ F/m, the vacuum permittivity. If at $x = 0$ (near the cathode) $V_0 = 0$ for simplicity, then $dV/dx = 0$. Physically this boundary condition says that owing to the space charge, i.e. the presence of ions all the way along the single charge trajectory, the initial acceleration is very weak or nil. With that, the Poisson equation may be integrated, and j found as a function of the accelerating voltage, V_{acc} , between the two grids at distance d using the Child-Langmuir law

$$j = \frac{4\epsilon_0}{9} \sqrt{\frac{2qV_{acc}^3}{d^4m_q}} \quad (11.25)$$

From Newton's third law, the thrust per unit area is

$$F = nm_q U^2 = j \frac{m_q}{q} U = j \sqrt{2V \frac{m_q}{q}} \quad (11.26)$$

Thus, the thrust scales with the square root of the voltage (of the potential energy).

11.6.2 Power

The ideal electric power is $P = jV$, which scales with $V^{5/2}$. Consequently, the power needed to produce thrust F is

$$\frac{P}{F} = \frac{I_{sp}}{2} \quad (11.27)$$

Commercial off-the-shelf GIEs have specific impulse of the order of 2,500–4,500 s. To produce a 1 N thrust requires about 25–40 kW in the practical case, because the

propellant is not 100 % ionized and losses (ionization, heating, and beam divergence, among others) raise the P/F ratio by a significant factor.

In space, electric power is typically gathered from solar panels. Even with gallium arsenide (GaAs) cells, the efficiency is only about 30 %, and since the solar constant in LEO is about $1,366 \text{ W/m}^2$ the solar panel surface grows rapidly with power. For instance, 10 kW needs about 25 m^2 . And recall that for deep space missions this number scales as the inverse of the square of the distance to the Sun. Note that improving the specific impulse lowers the propellant mass consumption but raises the power; which is proportional to the specific impulse cubed. Hence, each mission (characterized by payload, time, mass) must optimize the specific impulse and the thrust and that increasing the specific impulse is not the best strategy *per se*. For future deep space missions that require higher thrust, nuclear reactors (see Sect. 11.9) may power clusters of GIEs, freeing the spacecraft from dependence on solar power.

11.6.3 Propellants

The fact that the thrust scales with $\sqrt{(m_q/q)}$, the inverse of the specific impulse, implies a compromise in terms of the propellant molecular weight: everything being equal, heavier propellants (Hg, Xe, Cs) are better for thrust, but they will reduce the specific impulse. Mercury (Hg, $MW = 200$) was indeed the first propellant tested, but it has been abandoned due to its toxicity and the formation of amalgams when in contact with metals. The propellant almost universally used nowadays is xenon (Xe, $MW = 131$), extracted from air and expensive at about $4,000 \text{ US\$/m}^3$. Among the promising new propellants is bismuth (Bi, $MW = 208$), which must be vaporized but is non-corrosive. As for the voltage, the main constraints are to limit the power consumption and to avoid arcing. A compromise may be reached by having an initial grid to extract the ions, and then two grids, the first to accelerate, and the second, at a reduced voltage with respect to the first, to decelerate to some extent so that the net voltage will hold the power within the limitations of the solar panels.

11.6.4 Ion Optics (Gridding)

Older gridded ion engines (GIE) had only two grids, and their nominal voltage difference of 1 kV was increased to 5 kV to produce more thrust. However, such a high voltage results in high screen grid erosion. The most efficient grid arrangement thus far is four grids in two pairs. This method is based on the tokamak method of magnetic confinement.

Table 11.6 A range of some flown or space qualified gridded ion thrusters. NEXT and HiPEP were built at NASA-Glenn but never flown [5]

Thruster	Beam diameter (cm)	Specific Impulse (s)	Thrust (mN)	Power (kW)	Manufacturer
RIT-10	8.7	3,700	35	0.98	EADS
SERT II	15	4,770	29	0.91	NASA-Glenn
RIT-XT	21	6,419	218	8.06	EADS
T6	22	4,650	230	7.05	QinqiQ
XIPS-25	25	4,338	245	6.8	Boeing/L3-ETI
NSTAR	30	3,100	90	2.33	NASA-Glenn
NEXT	40	4,110	237	6.9	Engineering Model
HiPEP	91 × 41	9,620	670	39.3	Lab (JIMO Mission)

The first grid pair extracts ions by means of a moderate voltage of ~ 3 kV, with the first grid having a much larger open area ratio than the others. This eases the task of extracting and collimating the ions into a parallel beam. The accel grid potential is applied between the second and the third grids, separated by a wider gap. A negative potential between the third and the final fourth grid decelerates the beam, but also reduces the total power consumption.

After the beam leaves the thruster, it must be neutralized, otherwise the thruster and the spacecraft would be charged negatively by leftover electrons. The neutralizer ‘recycles’ electrons by injecting them into the ion beam outside the spacecraft, and closes the electric circuit. Injecting the electrons is typically done using a hollow cathode.

GIE are used for north/south station-keeping in geostationary satellites, but are also available for interplanetary probes. As their thrust is small, missions using GIEs have lifetimes of many years or even a decade. Thus, structural issues, not only electronics and electric issues are important. Among the structural issue associated with the grid, the most important are rigidity and erosion resistance. Grid spacing is of the order of 1 mm and there cannot be struts to stiffen the pairs of grids: these are rim held, and thus must be extremely rigid. This issue has limited the maximum diameter of a GIE to about 40–50 cm. Erosion due to ion impingement on the grid lattice, is addressed by the use of high Young’s modulus and refractory metals. Beryllium (Be) and molybdenum (Mo) have been used successfully. Carbon–carbon grids are an emerging technology because of their extreme rigidity, but are more prone to arcing due to protruding fibers in the holes. The issue of erosion is important in view of the required lifetimes of many years. Some performance data of GIEs are in Table 11.6.

11.6.5 Types of Gridded Ion Engine

Besides the gridding, ion thrusters are classified by their ionization technique. Kaufman’s GIE ionizes the propellant by bombarding it with electrons emitted by a hollow

cathode and directed towards an anode. Radio-frequency ionization does not need electrodes. It is produced by a periodic magnetic field that is excited by a coil powered by alternating current at a frequency of 10^3 – 10^4 Hz. These are called radio-frequency ion thrusters (RIT). The propellant may also be ionized by microwave heating. This technique does not require electrodes either. A detailed description of ion thrusters is given in [5].

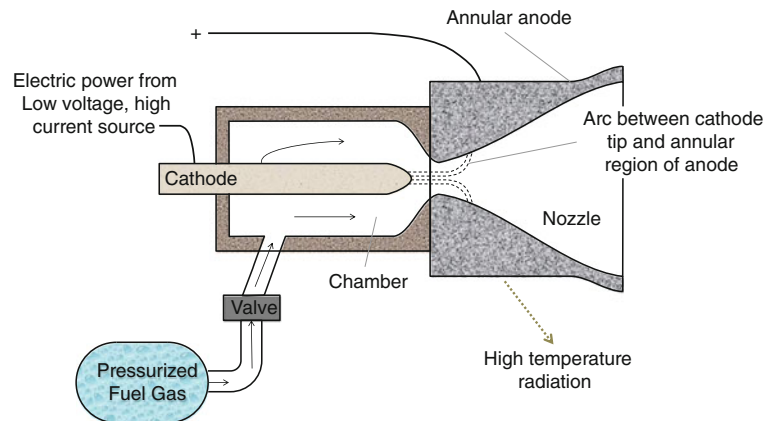
Ion thrusters tend to be reliable propulsion systems, but their thrust/unit volume is low compared to the Hall and MPD thrusters described in Sect. 11.8. Their thrust depends on power, and solar panels will probably be insufficient for certain classes of missions for example when the spent in space must be strictly limited due to radiation risks to a crew.

GIE technology has matured to the point of being commercial (e.g. the XIPS family of thrusters developed by Hughes), and is still evolving. The HiPEP thruster developed with NASA funding has demonstrated a specific impulse of about 19,000 s in the laboratory, at the price of a shortened life. The four-grid ion optic evolution has just started: ESA and the Australian National University tested the four-grid ion engine demonstrator DS4G in the laboratory, achieving a specific impulse of 210 km/s (about 21,000 s). Increasing the specific impulse depends on the accelerating voltage, V_{acc} , and there seems to be no conceptual obstacle to achieving specific impulses of the order 30,000–50,000 s, and even 100,000 s with the help of a tokamak or other magnetic confinement technique. At the same time, the electrical power, which scales with the cube of the specific impulse, will likely require a nuclear source.

11.7 Electrothermal Propulsion

Electrothermal propulsion was, in practice, one of the first attempts at electric propulsion. The principle is simply to convert electric power from, say, solar panels, into heat by using ohmic losses (resistivity losses), proportional to I^2R , with I being the current and R the resistance of the circuit.

Fig. 11.16 Schematic of an arcjet. Image Malcolm Macdonald



The energy released heats a propellant which is expanded in a conventional nozzle to produce thrust. This class of propulsion systems is called a resistojet. For instance, a 30 kW hydrogen resistojet can produce 6 N thrust at a specific impulse of about 860 s. Overall, the efficiency is of the order of 70 %. However, because hydrogen must be actively cooled for it to be space storable for periods of several weeks, other propellants must be used if a resistojet is to be used over longer periods, with an inevitable loss of specific impulse. The mass/power ratio of resistojets is much larger than for any chemical system.

Arcjets are a particular class of resistojet, where the conductor that dissipates electric power into heat is the plasma arc between the cathode tip and the anode, which typically consists of the conical nozzle itself. Thrust is produced by the expansion of the propellant gas, which is heated by the arc plasma as it flows coaxially to the cathode and through a throat, as sketched in Fig. 11.16.

The arc plasma may be stable ('anchor') even at a pressure of order 1 bar, and that sets arcjets apart from other electric thrusters, where the pressure must be much lower in order to prevent the ions and electrons from recombining. Thus, the most attractive feature of arcjets is their ability to work at significantly higher pressure than other thrusters, and to produce much higher thrust. In fact, the thrust per unit exit area may be several thousand N/m^2 , an order of magnitude higher than for current gridded ion engines.

As in all types of arc, the voltage is in the tens of volts, with a current that may reach 1,000 A. The longer the arc, the more power is converted into heat, so the arc should attach itself in the anodic diverging part of the nozzle and be spread as much as possible (not be attached to a single point), forming what is called a 'spoke'. When this can be realized, the arc attachment is said to be diffused. Although this desirable feature cannot be predictably controlled, increasing the propellant flow rate, \dot{m} , generally forces the attachment point to move downstream, with a desirable increase of current and power. Unlike chemical rockets, in recent arcjet designs the throat of the arc is shaped more

like a long duct with a constant cross-section. This constant cross-section duct is the 'constrictor' that forces more heat to be transferred from the arc plasma to the co-flowing propellant. In doing so, the voltage drop may be further increased and so may the power and the thrust. A second reason for the shape of the constrictor is the impossibility of predicting the exact position of the sonic throat, as the flow temperature increases from the cathode to the anode. Arcjet power is limited by the amount of energy that the arc can transfer to the propellant. This depends on the length of the arc, and therefore on its total resistance and voltage drop: the larger the drop, the larger is the power that is dissipated in the arc and transferred to the propellant, and the larger is the thrust. However, a constrictor with an excessive length to diameter ratio (L/D) also transfers more heat to the walls. The L/D ratio of the constrictor is therefore a compromise between these two effects.

Like a chemical rocket, the thrust and specific impulse of an arcjet scales with the ratio between temperature and the molecular weight (MW) of the gas propellant; i.e. on the $\sqrt{T/MW}$ factor. However, unlike a chemical rocket, no simple analytical relationship exists. It is experimentally known that the thrust scales with the electric power and inversely with the specific impulse (or exhaust velocity, v_e). A heuristic relationship defining the thrust efficiency, η , is the thrust power divided by the electric power

$$\eta = \frac{1}{2} \frac{\dot{m} v_e^2}{P} = v_e \frac{F}{2P} \quad (11.28)$$

showing that the thrust scales with the power via the efficiency, η , which is a quantity that must be experimentally determined. This efficiency depends on the geometry and architecture of the arcjet, and cannot typically be modeled with sufficient accuracy because the arcjet dissipates a large fraction (sometimes most) of the input electric power. The two main dissipation mechanisms are dissociation losses in the nozzle due to radical species that do not recombine in the nozzle and are ejected carrying their dissociation

energy, and radiative heat transfer from the hot arcjet body to space. An additional loss is ohmic heating of the electric circuit the feeds the thruster.

11.7.1 Dissociation Losses

These depend on the propellant gas, the operational parameters, and on the flow path through the thruster to the nozzle. Since arcjets are thermodynamic systems, to obtain a reasonably high specific impulse requires low molecular weight propellants. Thus, liquid or gaseous hydrides, for instance hydrogen (H_2), ammonia (NH_3), or hydrazine (N_2H_4), are preferred propellants, but helium and argon are also used. Molecular propellants however, dissociate promptly in the arc, forming many radical species such as NH , N , H and others, some of which are also vibrationally and electronically excited. As in a liquid rocket engine, these radicals recombine slowly in the divergent, and those that have no time to recombine represent a net energy loss that lowers the thrust and the specific impulse. At a specific impulse of the order of 2,000 s the bulk velocity is about g_0 times greater, at about 20 km/s. A nozzle diameter of the order of 0.5 m means a residence time of less than 1 ms, which is insufficient to recombine or neutralize all of the nitrogen, hydrogen or argon ions. In addition, the high temperature of the exhaust in the nozzle means a low density and a high viscosity. That is, along the nozzle the Reynolds number (Re) tends to decrease, and the radial velocity profile tends toward a distribution which resembles that of a laminar regime. An arcjet nozzle with an exit diameter of 10 cm and an area ratio of 100 may have an exit Re of the order of 100. Viscous effects dissipate much of the kinetic energy, and the expected performance does not follow the simple relationships in Sect. 11.2.

Note that, among the hydride propellants, hydrazine may also work as a monopropellant in the case of an electric malfunction of the arcjet. To this purpose, the propellant feed system must allow hydrazine to bypass the arcjet inlet and reach the hydrazine catalytic bed because hydrazine does not decompose spontaneously (see Sect. 11.3).

11.7.2 Radiative Losses

Radiative losses depend on how much of the arc heat is transferred to the walls, and hence to the external surface. Using the propellant itself as the coolant (regenerative cooling) recovers part of the heat loss, but the low specific heat capacity at constant pressure of common arcjet propellants cannot absorb much heat. There is an alternative advantage in regenerative cooling, and that is the increased pressure of the propellant injected in the arc chamber, which

increases the discharge voltage and the arc power. Data from [6] show that the simple radiatively cooled and the regeneratively cooled versions of the HIPARC arcjet tested at the University of Stuttgart had maximum efficiencies of about 30 and 37 %, respectively; in the second case that was found to depend also on the operational specific impulse, in fact falling with increasing specific impulse.

11.7.3 Technology

The major issues in arcjets are cathode and anode erosion, recombination losses, and power requirements.

Erosion is a common issue for all electric thrusters where ionization of the propellant is obtained using a thermal mechanism. Current must flow from the cathode to the anode by an applied voltage difference, which means that electrons extracted from the cathode will impact the anode. Ions will follow the opposite path. The extraction and impact are equivalent to energetically bombarding the electrode surface. Electrode surfaces may be damaged not only over the lifetime of the thruster, but also by a single improper ramping up of voltage and current during the startup of the arc, when the cathode is cold. The energy of the ions and electrons may be gauged from the voltage of the arc to be of the order of tens of eV, where 1 eV corresponds to 11,300 K in terms of temperature.

11.7.4 Power

High-thrust arcjets need substantial electric power. For instance, the ideal kinetic power, $\frac{1}{2}Fv_e$, to produce a thrust of 1 N is of the order of 10 kW, but that must be doubled or trebled to account for all losses. Because arcs work at 10–100 V, the current is typically of the order of thousands of A. Large amperage implies large ohmic (I^2R) losses that would be intolerable in space operation. To reduce them, the copper wiring must be massive. Future high-temperature superconducting wires, capable of remaining superconductive at liquid nitrogen temperature, may offer a solution to this problem, but their flight heritage, although flown on-board the TECHSAT II in 1998, is limited.

11.7.5 Performance

Arcjets at 100 kW are capable of thrusts of the order of 1–5 N, and specific impulses that vary between 1,000 and 2,000 s. In general, the specific impulse grows with the thermal load of the arcjet, but the efficiency is typically less than 50 %.

11.7.5.1 Materials Technology

The arc plasma temperature may vary locally from 15,000 to 25,000 K, and the arcjet body, although shielded by propellant, must withstand temperatures of 1,000 K and higher. Thus, the best body materials are refractory alloys or ceramics such as boron nitride (BN, melting point 3,246 K), tungsten (W, melting point 3,410 °C), hafnium oxide (HfO₂, melting point 2,758 °C) and others, all of which are costly and expensive to work with. The choice of material depends also on the electric properties and electronic extraction work (the ‘work function’ is ~4–5 eV for metals). In order for electrons to be steadily emitted from the tip of the cathode, this requires the cathode tip temperature to be maintained at over 2,000 K. This requirement is not satisfied at startup, and if the voltage ramping is not right the cathode may locally melt, resulting in fast erosion, pitting, and unstable operation. Chemical attack may also start if the propellant contains oxidizing impurities, including water or its traces. There is no standard remedy for this malfunction, except to size the cathode so that its surface area is large enough to distribute the heat flux evenly, and to choose a ‘good’ refractory material. To reduce the electronic extraction work, electrodes are frequently made with a dispersion of thorium oxide (‘thoriated’ electrodes), but this adds to the manufacturing cost.

In-space cooling of arcjets is typically by radiation, which carries an efficiency penalty. Regenerative cooling increases efficiency, as noted, but adds to the material manufacturing cost because channels to extract heat from the arcjet body and preheat the propellant must be milled, and refractory materials are notoriously very hard and more fragile than metals, making them more difficult to work with.

11.7.5.2 Testing

This is a serious issue for high-power arcjets (in the hundreds of kW). In practice this requires a dedicated power station attached to the laboratory, complemented by a power conditioning unit to transform AC current from the high voltage grid to the low DC voltage required by the arc. This adds further to the cost and complication of testing a high power arcjet: in fact, simulating arcjet performance in the vacuum of space needs a vacuum tank and significant pumping power, since the specific impulse is of the order of 2,000 s at most, and the mass flow rate is at least 1 g/s. Not many facilities exist worldwide that are capable of this performance; one is at NASA-JPL and it is actually capable of handling several MW of power.

In conclusion, arcjets are a class of thruster that is capable of producing a thrust of several newtons, which is much higher than most electric thrusters. They are also compact and relatively uncomplicated. Their disadvantages are that the specific impulse is typically limited to

1,200–1,800 s, and the low efficiency raises thermal issues in disposing of the waste heat. These disadvantages weigh heavily on the evolution of high-power arcjets. Consequently, at this time they are not in common usage in space. They could perhaps enjoy a second life when space nuclear power becomes available.

11.8 Magnetoplasmadynamic Thrusters

Magnetoplasmadynamic (MPD) thrusters work by accelerating ionized matter by the Lorentz force, $\mathbf{F} = \mathbf{J} \times \mathbf{B}$, where \mathbf{J} is the current flux (coulombs per unit area and per unit time, or A/m²) produced by ion movement, and \mathbf{B} is the magnetic induction vector (in tesla, T). The vector product implies that the force applied to the ion stream is normal to the plane in which \mathbf{J} and \mathbf{B} lie. Thus, for the Lorentz force to exist, a current (ionized gas, for instance) must pass between an anode and a cathode, and there must exist an electric field and a voltage driving the current. Once the ions move, driven by this voltage, the Lorentz force of the magnetic field \mathbf{B} will accelerate the ions. The reaction (thrust) is the Lorentz force applied to the circuitry necessary to maintain the voltage and the magnetic field.

Unlike a chemical thruster, where the thrust direction is fixed by the nozzle axis, the thrust of an MPD thruster depend on the particular configuration of the electric and magnetic fields, and of their electric circuitry. There may also be transients, e.g. a magnetic field can be generated by the current flowing instantaneously between electrodes. The electric and magnetic fields are degrees of freedom in conceptually designing an MPD thruster. There is no general theory of MPD thrusters: their performance (thrust and specific impulse) cannot be estimated based on knowledge of, say, the applied voltage and magnetic field. In fact, in pulsed thrusters, *no* magnetic field is applied, it is created by the current driven by an instantaneous voltage. Although thrusters are often specified in terms of electric power, that is the energy consumed, not the thrust delivered.

The simplest MPD thruster is shown in Fig. 11.17. A spark vaporizes and ionizes a minute amount of propellant, here Teflon, fed to the ionization chamber by a loaded spring. Ionized Teflon products driven by the voltage form an instantaneous current, I , between the electrodes. This current produces a magnetic field \mathbf{B} normal to the plane of the figure. \mathbf{B} is also normal to I , and the $\mathbf{J} \times \mathbf{B}$ Lorentz force on the Teflon plasma accelerates it from left to right. When the plasma slug is ejected, the thruster stops, and a new spark must resume the process. This MPD thruster is called a *pulsed* plasma thruster (PPT), and it can be practically realized in many ways. For instance, Fig. 11.18 shows an axisymmetric PPT in which the current itself creates the magnetic field whose azimuthal component, B_θ , is normal to

Fig. 11.17 Schematic of a pulsed plasma thruster. *Image U. Walach*

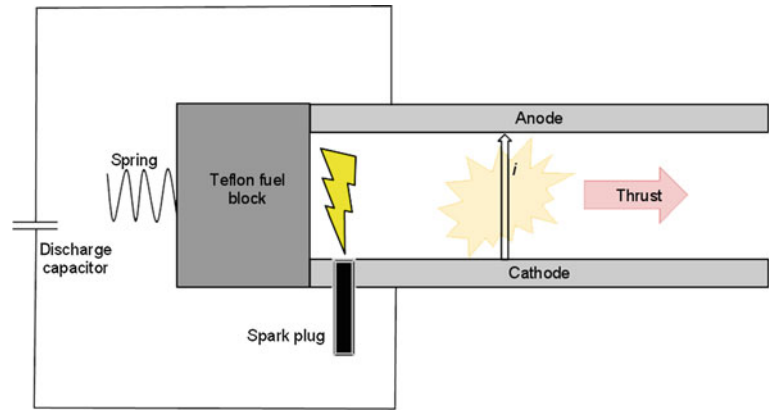
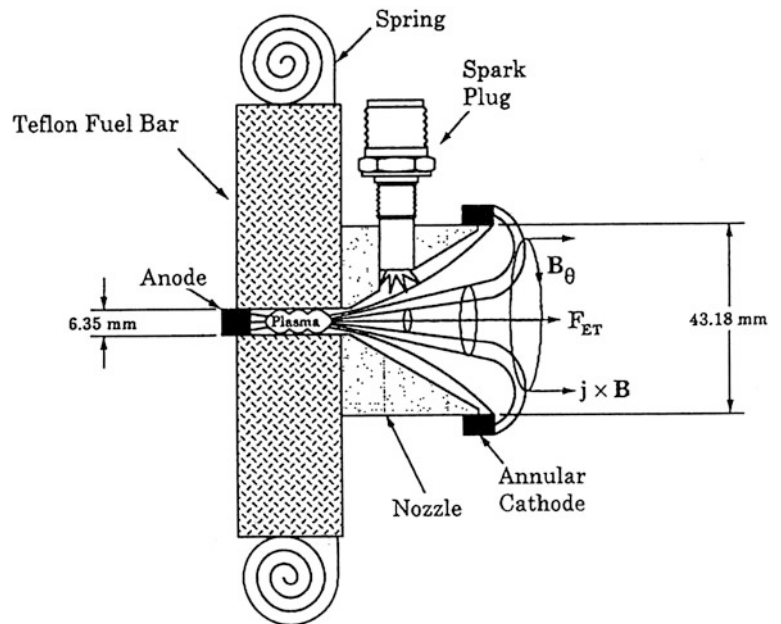


Fig. 11.18 Schematic of an axisymmetric pulsed plasma thruster. *Image University of Michigan, Non-equilibrium Gas & Plasma Dynamics Laboratory*



the current in most of the nozzle and serve to accelerate the plasma.

The instantaneous magnetic field created by a pulsed thruster may be large, but the current depends on the quantity of material that is ablated and ionized to form the current. Most PPTs are small, e.g. in the hundreds of μN to a few mN range, but their specific impulses may reach 2,000 s, although with a low efficiency of perhaps 10 %. A characteristic quantity is the impulse bit, the product of thrust times time. Small PPTs for satellite attitude control can produce impulse bits (Ns) in the range 100–1,000 μNs , consuming several tens of Watts. A micro-PPT built at the University of Washington, powered by 12.5 W from solar panels, had a maximum impulsive thrust of 0.14 mN, a specific impulse of 500 s, and an impulse bit of 70 μNs ; the energy consumption/ ΔV was $1.4 \times 10^5 \text{ J}/(\text{m/s})$ for a total mass of 3.8 kg. PPTs are used on microsattellites because, although not very efficient, they are simple, reliable, easy to

operate, and with solid propellant can last for many years without maintenance.

PPTs are normally self-field MPD thrusters, i.e. they generate the magnetic field \mathbf{B} with their own current. Self-field thrusters may also produce continuous thrust. A second class has the magnetic field applied externally, and this is the technology of choice for producing much higher continuous or quasi-continuous thrust. The simplest of these is the Faraday thruster, or Lorentz Force Accelerator (LFA). Conceptually, it consists of a square channel with walls that are insulated from each other. Applying a voltage of several hundred V between two parallel walls creates an electric field, \mathbf{E} , and ionized propellant driven by the electric field establishes a current between the two walls. If the power available is in the hundreds of kW, then the current may be of order 1–10 kA. The applied magnetic field can be created either by permanent magnets or by electromagnets, the two poles formed by the pair of parallel walls. In principle, this

Fig. 11.19 An MPD thruster with a magnetic nozzle. *Image credit* the High-Power Electric Propulsion Laboratory (HPEPL) in the Georgia Institute of Technology Department of Aerospace Engineering



arrangement maximizes the Lorentz force accelerating the plasma, since \mathbf{J} and \mathbf{B} are at right angles to each other. Thrusters may be axisymmetric, except that a solenoid should be imagined wound around the divergent (conical) nozzle. The solenoid creates an additional \mathbf{B} field that helps to raise the Lorentz force and to stabilize the plasma. Such an MPD thruster and nozzle arrangement is shown in Fig. 11.19. This simple description belies the technological difficulties of handling large currents. With xenon (Xe) and argon (Ar) gases the efficiency is 20 % at most, but lithium (Li) and dihydrogen (H_2) it can be as much as 50 %. Overheating and erosion of the electrodes (especially the cathode) is common. MPD thrusters are capable of absorbing power in the MW range, but their low efficiency requires the unused power to be radiated away to space, either directly from the engine or, less preferably, through a space radiator.

Axisymmetric self-field and steady MPD thrusters resemble arcjets, except their power may be much higher. Steady MPD thrusters tested in Russia at the Moscow Aviation Institute have been operated at 188 kW, producing specific impulses of the order of 4,500 s with 49 % efficiency. A 250 kW and a 500 kW radiatively cooled MPD were designed and lab tested at NASA-JPL with an efficiency about 60 % and specific impulses of 4,500 and 6,200 s, respectively. In general, such power levels are hard to test in the laboratory; space operation at power over 1 MW becomes impractical with solar panels and requires nuclear power. Experience with MPD thrusters in the 1-MW class shows that lifetime is inversely correlated to current. Using lithium propellant and adding a small percentage of barium (Ba) increased the lifetime from 8,800 h at 10 kA to 28,000 h at 2.75 kA. Thrust and specific impulse depend on the combination of power and propellant. Lithium (with barium additive) yields a maximum power 1–5 MW, with a thrust of the order of 1–2 N/MW and a specific impulse of 4,000–6,000 s. Above 5 MW only

hydrogen can absorb the voltage drop and power at an efficiency >60 %, and the specific impulse may rise to 10,000–15,000 s because of the low molecular weight. Obtaining sustained performance at this level requires long-life cathodes, for instance perforated and internally cooled, and a high emissivity coating such as zirconium diboride (ZrB_2).

The trend toward powers of many MW depends on the fact that in most designs the efficiency increases with the power. Several pulsed MPD thrusters have been designed (but not tested) that are capable of absorbing a few MW. However, their efficiency is still below 50–60 %, not only because of ionization losses but also because the ion current ejected (i.e. the ion beam) is not collimated well. In fact, to improve the thrust direction, the applied field must be made into a true magnetic nozzle. Alternatively, the geometry should be that of an LFA. The thrust density and power/volume of multi-MW MPD thrusters are about an order of magnitude higher than with gridded ion engines. Once the erosion problem is solved, MPD systems may become competitive for space missions.

11.8.1 Hall Thrusters

Hall thrusters were first developed in the former Soviet Union in the 1960s by the Fakel design bureau, and their technology has been exported to France and to the US. More than a hundred Hall thrusters have been space flown by the Soviet Union and Russia in the past 40 years.

The principle of the Hall thruster is shown in Fig. 11.20. It is based on the difference in mass between ions and electrons in a magnetic field, which determines the order of magnitude differences in gyration radii and frequencies. Electrons extracted from an external cathode move toward the internal anode, are captured by an applied radial magnetic field, \mathbf{B} , and drift in the channel, hence the technical

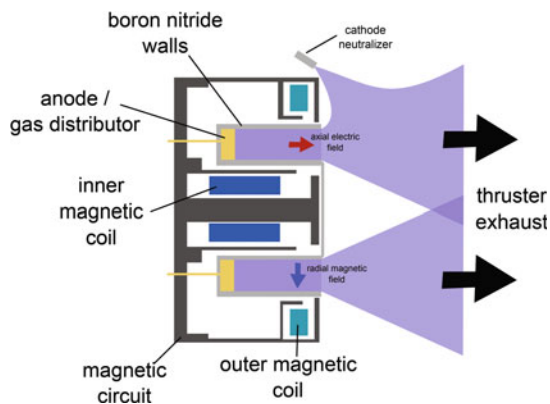


Fig. 11.20 Schematic cross-section of a radially-symmetric Hall effect thruster. Electrons emitted by the cathode are trapped by the magnetic field, and act as a virtual cathode. Ions, created by the collision of neutral propellant with electron are accelerated from the anode to the virtual cathode, acquiring momentum that prevents them from recombining with electrons. The electrons emitted by the external cathode neutralize the ion beam. *Image* Finlay M^cWalter

name of closed electron drift thrusters. Propellant (xenon, Xe) is injected, collides with electrons and ionizes, and the ions move along the electric field, \mathbf{E} , created by the voltage drop in the magnetized plasma.

In a magnetized plasma, the resistivity perpendicular to the magnetic field lines is much higher than along the lines. Therefore, the electrostatic field is perpendicular to the magnetic field lines. The ions do not recombine with electrons because their momentum is approximately 10^4 times larger than that of the electrons, and they are accelerated out. This produces a thrust in the 0.1–1 N range, with the latter having been demonstrated in experiments but not yet flown. The efficiency is about 50–60 % and the specific impulse is 1,500–3,000 s. The discharge voltage is from hundreds of V to 1 kV. The power of the Hall thrusters flown so far is in the 10 kW range, and their erosion is lower than for self-field or applied-field MPD thrusters. The PPS[®] 1350 thruster has been qualified at 10,500 h, which is sufficient for some interplanetary missions, and it was used on the ESA SMART-1 probe that spiraled out to the Moon and entered orbit there.

11.8.2 VASIMR Thruster

The VASIMR (Variable Specific Impulse Magnetic Rocket) thruster, developed initially by Franklin Chang-Díaz (born 1950), at NASA, and now by the Ad Astra Company in Texas and Costa Rica, differs from other MPD thrusters in its ionization strategy and thrust control. The ionization is achieved by radio-frequency (RF) electromagnetic waves of 10^3 – 10^4 Hz, as in tokamak magnetic confinement technology. A helicon antenna broadcasts RF energy to the

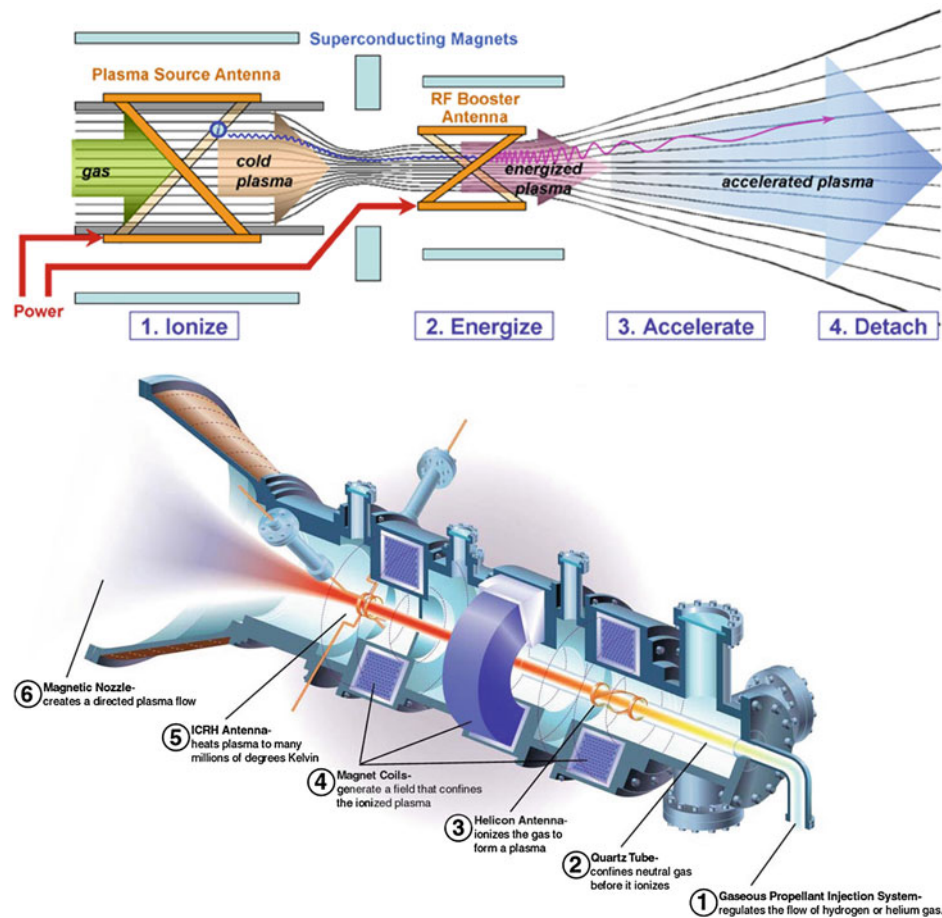
propellant molecules, increasing their internal energy degrees of freedom, as a preliminary to ionizing them via collisions. This strategy eschews electrodes entirely, and thus the erosion problems that limit all other electric thrusters. The second stage of VASIMR uses RF matching the ion cyclotron frequency to further increase the energy of the ions. Before the ions can redistribute their translational energy into vibration and rotational degrees of freedom, a magnetic nozzle ejects them. Thus the magnetic field, \mathbf{B} , in VASIMR plays the key role in the ion acceleration process, see Fig. 11.21. The magnetic field also plays a key role in preventing the extremely energetic plasma from melting the thruster wall, because it traps ions within its lines of force. As the magnetic field controls ionization and acceleration, VASIMR should be able to vary its thrust against specific impulse on-demand, for instance using a high thrust and a low specific impulse when spiraling out from LEO, and increasing specific impulse and lowering the thrust when traveling along a interplanetary trajectory.

VASIMR prototypes are labeled with their VF-number, indicating the power in kW. Their thrust efficiency has grown from 40 % for VX-50 to an estimated 50 % for VF-200, which is due to be tested on the International Space Station (ISS) post-2015. The power that does not contribute to kinetic energy ends up as heat that must be disposed of in some way, for instance by radiation. The magnetic field to operate VASIMR is of the order of 1–2 T, which requires cooling conventional electromagnets, or cryocooling of superconductive solenoids. Since VF-50 was capable of 0.5 N, assuming 50 % efficiency and specific impulse of 5,000 s, the thrust of the VF-200 should be about 2 N. VF-200 will be tested in an experiment where it will temporarily replace the conventional MMH/NTO thrusters that are normally used to reboost the ISS orbit, which is constantly decaying due to atmospheric drag.

11.8.3 Propellants and Magnetoplasmadynamic Life

The technology of MPD thrusters is not as well established as that of ion (electrostatic) thrusters, although trends in terms of power, propellant and materials are beginning to be understood. Xenon, argon, H_2 and its isotopes, and recently lithium and barium, have all been tested. The light elements (H_2 , He and Li) give the best specific impulse performance for applied-field MPD, but H_2 and other gases erode the electrodes. As ionization losses are important, and it is recognized that the heavy molecules that are traditionally favored as propellants (Xe, Ar) may dissipate part of the input power in ionizing the L-shell because the second ionization potential is higher than the first. In fact, lithium has a second ionization potential that is so high as to be

Fig. 11.21 Principle of operation of the VF-200 VASIMR prototype showing the diverging ion trajectories at the nozzle exit (*top*), and schematic VASIMR layout (*bottom*). Image Ad Astra Corporation



‘inaccessible’, and this reduces losses. The performance of steady MPD thrusters varies with the current, the applied magnetic field, and the size. Thrusts up to several newtons and specific impulses up to 5,000 s have been obtained in the laboratory with light propellants (Li and H₂ or D₂) at efficiencies up to 50%. Heavier gases like Xe and Ar produce specific impulses near 2,500 s. The erosion of electrodes with most gaseous propellants limits the life of all MPD thrusters, but lithium propellant seems to improve the electrode lifetime considerably, and especially when barium vapor is also present. The power that can be absorbed by an MPD thruster depends on the ionization losses and on the terminal voltage drop. It also depends on the propellant. Lithium propellant can work up to a drop of several hundreds of volts, in practice limiting the power to several MW because currents above 4–5 kA erode the cathode too rapidly. To absorb more power a voltage drop of many hundreds of V is required, and because hydrogen and its isotopes can do it this is the most likely propellant of choice for future thrusters in the tens of MW. An additional benefit of hydrogen is its specific impulse, which should rise above 10,000 s.

The lifetimes of MPD thrusters can typically be measured in weeks of operation, which is inadequate for

interplanetary missions where their relatively high thrust would be very useful. Hall thrusters have a life of the order of many months to a year, but their specific impulse is at most 1,500–3,000 s. VASIMR works best with hydrogen and deuterium and, being electrode-less, should in principle be capable of a much longer life than gridded ion engines and MPD thrusters. Cathode technology using refractory materials, active cooling, and high thermal emissivity coatings, is progressing and may make future MPDs, powered by nuclear reactors, the propulsion technology of choice for crewed missions.

11.9 Advanced Propulsion

Advanced propulsion includes many innovative and conceptual systems that do not fall within the conventional categories. For instance, nuclear propulsion (all types, such as fission, fusion, and matter–antimatter annihilation); interstellar ramjets; beamed energy propulsion; solar and magnetic sails that use the photons and protons emitted by the Sun; and even more exotic systems that are based on general relativity and quantum mechanics, such as space-warping and zero-point energy.

Most of these are either just concepts, or they have been proven to be realizable only in the very far future; as for instance, fusion propulsion. Some have been proved to be impractical engineering-wise, such as beamed energy for interplanetary travel, or flawed in the sense they violate currently understood fundamental physics, like interstellar ramjets. Among those that have been at least ground-tested, there is nuclear propulsion, an old technology born after WWII, and solar sails, which have been actively pursued by several space agencies since the 1960s and recently demonstrated in space.

11.9.1 Nuclear Propulsion

Why nuclear propulsion? The scaling of specific impulse, thrust, and power offers the answer.

- Specific impulse and the propellants mass flow rate, \dot{m} , scale with the exhaust velocity, v_e , and the specific impulse controls the total mass consumed.
- Thrust, F , scales with the mass flow rate, \dot{m} , times the exhaust velocity, v_e , and hence with v_e^2 it controls the mission duration.
- Power is the kinetic energy per unit time ejected, or $\frac{1}{2}\dot{m}v_e^2$. It scales with v_e^3 , and is the price to pay.

Due to the radiation doses to any crew, deep space missions must be faster than can be achieved using chemical propulsion. Thrust must be applied for much longer than the few minutes of a chemical rocket. The energy consumed will be far greater than that available using chemistry. For instance, a thrust of 1 N by a future ion engine with specific impulse 10,000 s needs 0.1 GW, and since the thrust is only 1 N, this power will have to be maintained for a long time. The potential energy that must be available to do that depends on the fundamental forces of physics, see Table 11.1.

The ideal energy balance, energy density equals $\frac{1}{2}v_e^2$ shows that the exhaust velocity, v_e , equals the square-root of twice the energy density. To increase the thrust and reduce the travel time the exhaust velocity must be made as large as possible. Note that by doing so the mass consumption decreases but the price is a dramatically increased by the on-board power that must be available.

Table 11.1 shows that the nuclear force is the only one that can meet the challenge posed by future missions, in particular human missions to Mars, the asteroids, or the icy moons of Jupiter. Any other ‘slower’ solution, including solar, appears unfeasible in view of the fact that the radiation dose outside the van Allen belts is in the Sievert range for a 1-year mission (a yearly dose on Earth is 0.6–2.4 mSv, depending on location). It is generally agreed that human missions to even the closest planets must be

fast (3–4 months, at most) and that only nuclear propulsion can provide the required energy [7].

The main strategies to exploit nuclear energy are thermal, electric, and a combination of the two.

11.9.1.1 Nuclear Thermal Propulsion

Nuclear propulsion (NP) was originally proposed at the end of WWII in the UK, in the wake of the atomic bomb tests. The US Department of Defense (DoD) started investigating nuclear propulsion for the second stage of the Atlas intercontinental ballistic missile (ICBM) under the Rover project. After the Atlas problems were solved, ROVER became NERVA (Nuclear Engine for Rocket Vehicle Application), and the application was space flight. The Soviet Union’s history of nuclear propulsion remains unclear; however, it very likely followed a technology path similar to that in the US.

Nuclear thermal propulsion (NTP) involves the use of compact nuclear reactors (NR). The fission of the uranium-235 (^{235}U) isotopes releases fission fragments, neutrons, and gamma rays which by colliding with nuclei of uranium and others species, convert (‘thermalize’) their kinetic energy into heat. The kinetic energy is up to 200 MeV for fission fragments, and 10 MeV for neutrons and gamma rays. By means of cooling channels manufactured inside the fuel bars, this heat is transferred to the coolant (LH_2), which is gasified by the heat and is expanded in a conventional nozzle. This type of reactor, where the fuel is contained inside replaceable bars, is called a solid-core nuclear reactor. Other core concepts include liquid and gas-core reactors, and these were investigated at the Kurchatov Institute in Russia and at the Los Alamos Scientific Laboratory (LASL) in the US.

LASL designed and tested solid-core reactors for NERVA, and Westinghouse/Aerojet engineered them into actual rocket engines and tested them, see Fig. 11.22. Most of the LASL work was to design the fuel bars and their cooling channels so that their lifetime and integrity could be ensured for at least several hours. The fuel was ^{238}U enriched with fissile ^{235}U . As in commercial reactors, the fuel was clad by appropriate materials, e.g. zirconium alloys. To enable operation at high temperatures, the low melting point (1,405 K) uranium had to be replaced by its UO_2 oxide (melting point 3,138 K). Nitrides and carbides of uranium were also tested successfully up to 2,775 K. The most powerful Los Alamos reactor was Phoebus IIA, tested at 4.2 GW for 12 min. With a nozzle, the specific impulse would have been 890 s, and the thrust approximately 900 kN. Less powerful (10–100 MW range) but more versatile nuclear reactors and rocket engines were also designed and tested, see [8]. In the Soviet Union, ternary alloys were allegedly tested to 3,275 K.

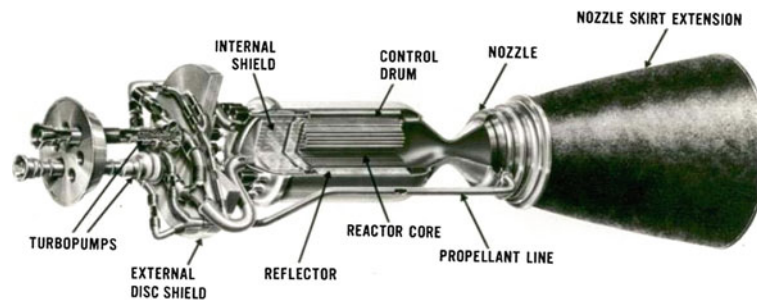


Fig. 11.22 Westinghouse-Aerojet NRX nuclear thermal space rocket (1965). The control drum absorbs neutrons depending on rotation

angle in order to control the thrust, which was 75,000 lbf (333.6 kN) at specific impulse of about 880 s

The United States Air Force (USAF) took over nuclear reactor research in 1972, when NERVA was canceled. It developed ceramic fuel (CERMET) reactors in the 1–2 GW range capable of withstanding hundreds of on–off cycles without cracking, and very compact nuclear reactors using fuel pebbles instead of fuel bars. These were called pebble bed reactors (PBR). These proved to be prone to thermo-hydraulic instabilities. The Brookhaven National Laboratory (BNL), working for the US Navy, produced a PBR that could be packaged inside a standard-size torpedo, about 25 inches (635 mm) diameter. Its power was in the tens of MW range, but the reactor life expectancy was only a few minutes. The thermal power/unit volume was of the order of 1 MW/liter. This technology has been proposed for in-space propulsion by the Plus Ultra Technologies company in the US under the acronym MITEE. Plus Ultra has also investigated very compact nuclear reactors using low critical mass fuels, for example, americium-242. Similar work was done in Israel at Ben Gurion University, and his work has produced 1-m size reactors capable of several tens of MW (just as for MITEE). Properly engineered, this reactor was proposed for a crewed Mars mission. This nuclear thermal propulsion engine is similar to one independently proposed by C. Rubbia in the 1990s.

Liquid- and gas-core reactors have been investigated in Russia and in the US, but never built. Their theoretical specific impulse may be expected to be at least 1,500 and 3,000–5,000 s, respectively.

11.9.1.2 Nuclear Electric Propulsion

The second nuclear propulsion strategy uses the reactor to generate electric power to run one or more of the thrusters described in Sects. 11.6–11.8. This is called nuclear electric propulsion (NEP). The basic components are the reactor, a prime mover using a Rankine, Brayton, Stirling or other cycle (for instance a gas turbine) powered by the working fluid of the cycle, an electric generator driven by the prime mover, and the electric thruster. The thermodynamic cycle may have an overall efficiency of the order of 30–40 % and

the remainder of the reactor thermal power must be dumped. In space, this means a space radiator. A space radiator is the most cumbersome and weighty element of a nuclear electric propulsion system, sometimes weighing more than the reactor itself. The power radiated per unit mass is of the order of 0.4–1 kg/kW for metallic radiators but much less for carbon nanotubes structures.

Gridded ion engines are already mature for nuclear electric propulsion, but due to their low-pressure operation, their thrust is very small, at most 0.1–0.5 N for the foreseeable future. Their specific impulse may however reach hundreds of km/s. Hall thrusters have already demonstrated 1 N in ground testing, and can deliver specific impulses in the 2,000–3,000 s range (up to 5,000 s with lighter propellants) and lifetimes in excess of 10,000 h. MPD and arcjets are capable of higher thrust, but have not yet demonstrated specific impulses beyond 5,000 s, and must mature further.

The low thrust of all electric thrusters means that, for significant spacecraft acceleration and reduced mission time, either the engine must be ‘on’ for most of the mission, or that the nuclear electric propulsion system must consist of tens or hundreds of thruster modules, or both. The first approach may be acceptable for scientific missions but poses severe reliability issues, and because the ground team may be active for years, will be costly. A key issue with years of operation of high power gridded ion thrusters is the life of the grid. The high voltage means larger holes, but thicker grids can be used to offset some of the problems. With most MPD thrusters that use electrodes, the key issue is anode/cathode erosion. And of course the nuclear reactors must be capable of operating reliably and autonomously for years. One problem is reactor refueling, if the life of the rods, bars, pins or pebbles is shorter than mission time. The second approach of increasing the total thrust by having multiple thrusters, looks more practical, but it requires plenty of electric power. However, it increases overall safety and reliability, because mission times may be shortened to months instead of years.

11.9.1.3 Dual-Mode Nuclear Propulsion

Dual-mode propulsion seeks to optimize missions by combining the best of the two nuclear propulsion strategies. Nuclear thermal propulsion produces thrust similar to chemical propulsion, but the specific impulse is typically less than 1,000 s. Nuclear electric propulsion is mass-thrifty, but its thrust is inadequate, except where multiple thrusters are used. Russian work emphasizes the bi-modal approach, also proposed by NASA-Glenn, whereby part of the nuclear reactor energy is for thrust, part is to power the systems needed for the crew, and part is fed to the electric thrusters. Using a crewed Mars mission as a scenario, nuclear thermal propulsion allows escape velocities from LEO to be reached in just a few revolutions of a spiral trajectory. Then the nuclear thermal propulsion is switched off and an electric thruster (with specific impulse 4,000–15,000 s, and thrust 1–10 N) is turned on to power the interplanetary leg of the trajectory and to decelerate. The nuclear thermal propulsion is used again for Mars orbit capture. Note that the same propellant (H_2) can be used for both propulsion systems. Dual mode operation avoids the very long spiral trajectories imposed by the low thrust typical of all electric thrusters, and thus avoids transiting the deadly van Allen belts many tens or hundreds of times while accelerating to escape velocity.

Plus Ultra Technologies has proposed a similar dual-mode concept based on MITEE [9]. The VASIMR engine in development by the Ad Astra Company is an ‘infinitely variable-mode’ system: in principle, it is capable of varying both its thrust and specific impulse so that their product (power) stays constant. The specific impulse of the VASIMR prototypes is about 5,000 s, and while its thrust is small it’s at least ten if not a hundred times larger than that of a gridded ion thruster. As it is electrode-less, it may be more reliable than high power arcjets or most MPD systems, although at the cost of more complexity (RF generators, superconducting magnets, magnetic nozzle).

11.9.1.4 A Comparison of Nuclear Electric Propulsion Versus Nuclear Thermal Propulsion

NERVA-style nuclear thermal propulsion is capable of producing an extremely large thrust at a specific impulse in the 880–900 s range. This technology is costly but feasible, and with PBR and higher temperature ceramics the specific impulse may reach over 1,000 s. Nuclear electric propulsion consumes less propellant, but at the price of low thrust. An important issue with nuclear electric propulsion is converting thermal power to electricity. A gas turbine Brayton cycle is best, if the gas turbine and its turbo-alternator or dynamo can work unattended for months on end. This gas turbine technology ‘package’ is now mature in the 1 megawatt thermal (MWth) range. A conversion efficiency,

η , of approximately 25–30 % is feasible, with a total nuclear electric propulsion system mass of the order of 1–2 t. Heat extraction may be based on a gas cycle, a liquid metal cycle, or require heat pipes. In this context, the third drawback of nuclear electric propulsion is its ‘bottom’ temperature, at which the $(1 - \eta)$ of thermal power that is *not* converted to mechanical work must be dumped by the space radiator because if it is too low the radiator cannot radiate away enough waste power and if it is too high the efficiency drops.

Gridded ion engines are evolving towards 15,000 s, and potentially towards even higher specific impulses by exploiting magnetic confinement technology, such as tokamak; MPD thrusters may evolve rapidly if the electrode life can be extended, and their eventual power and thrust may be much higher. Their power/thrust ratio is still in the 40–60 kW/N range, thus the power/mass of nuclear electric propulsion is about a few percentage points of nuclear thermal propulsion.

Despite the stated disadvantages of nuclear electric propulsion, its advantages are remarkable in terms of integration, e.g. in coupling turbine, electric generator and thruster(s): all of which can be optimized separately. Europe and the US have extensive gridded ion engine technology, while Russia leads in MPD and Hall thrusters. In addition, Russia has conceptually designed many dual-mode engines, and advocates nuclear propulsion as the best solution for crewed Mars missions.

11.9.1.5 Pulsed Nuclear Thermal Propulsion

This is a conceptual form of nuclear thermal propulsion employing nuclear explosives (bombs). Developed initially from an idea by S. Ulam in 1958, it became Project Orion, funded in the US DoD for several years. It consists of detonating a sequence of atomic bombs ejected astern of a spacecraft, at a distance precisely matching the size of the fireball and the spacecraft’s pressure plate. The expanding gas from each explosion has a velocity that depends on the yield. For instance, a 1-kt explosion has an ideal expansion velocity of about 500 km/s, producing a very large dynamic pressure on impact. The spacecraft receives each momentum/pressure pulse on its pressure plate, a flat structure that is connected to the spacecraft by shock absorbers. These must be dimensioned to reduce the impulsive acceleration to a few g_0 . The original concept envisioned ejecting a series of canisters through a hatch in the center of the plate, each containing a bomb and some propellant to enhance the gas expansion. The explosion frequency needed was estimated to be a few Hz.

Calculations predicted specific impulses linearly proportional to, v_e , with v_e being the velocity of expansion of the fireball. By optimizing the yield and the distance, the constant of proportionality was found to be in the range

0.1–0.5. This concept was successfully tested using conventional explosive, however Project Orion was terminated in 1963.

Its recent successors use a magnetic field to compress and fission Curium-245 (^{245}Cm) fuel. The ionization and magnetic fields created by the explosion produce pulsed Lorentz forces in a magnetic nozzle. This concept is the so-called Magnetic Orion (MagOrion) project, investigated for the USAF by Andrews Space in 2000 and 2003. Its simpler and smaller version is Project MiniMagOrion. The specific impulse of MiniMagOrion was estimated to be 10,000 s, sufficient to enable a 100-t spacecraft to reach Mars in three months. However, the cost of validating this concept, its complexity, and doubts about the coupling efficiency between the electromagnetic pulse and the magnetic nozzle halted the project in 2003.

11.9.2 Solar Sails

Harnessing the power of the Sun to propel a spacecraft may appear somewhat ambitious and the observation that light exerts a force contradicts our everyday experiences. However, it is an accepted phenomenon that the quantum packets of energy which compose sunlight, that is to say photons, perturb spacecraft through the conservation of momentum; this perturbation is known as solar radiation pressure (SRP). To be exact, the electromagnetic energy from the Sun pushes the spacecraft and Newton's second law states that momentum is transferred when the energy strikes and when it is reflected. The concept of solar sailing exploits SRP to propel a spacecraft, potentially providing a continuous acceleration limited only by the lifetime of the sail materials in the space environment. The momentum carried by an individual photon is extremely small. At best, a solar sail will experience a force of 9 N/m^2 of sail at Earth's distance from the Sun, so to provide a suitably large momentum transfer the sail is required to have a large surface area, while maintaining as low a mass as possible. Adding the impulse due to incident and reflected photons reveals that the idealized thrust vector is directed normal to the surface of the sail. Hence by controlling the orientation of the sail relative to the Sun the spacecraft can gain or lose orbital angular momentum. Using the momentum gained by reflecting these quantum packets of energy, the sail can slowly but continuously accelerate to accomplish a wide-range of potential missions [10].

One of the key problems in solar sail design is the packing and subsequent deployment of a large area of thin film. The dimensional expansion ratio between a deployed and stowed solar sail can be over 100, thus innovative structural engineering solutions are required. The packing and deployment problem has perhaps been one of the

greatest impediments to practical solar sail utilization. In addition, since the sail is folded for packing, the reflecting medium of the sail must be mounted on a thin substrate. The presence of a substrate imposes a fundamental limitation on the performance of the sail due to the parasitic mass, this being defined as the total non-reflective mass of the solar sail and attached spacecraft. The conventional belief is that the solar sail film must be as flat as possible in order to maintain as high a reflectivity as possible. If this is true, then tension must be applied to the deployed sail, either by a deployable structure or by spin-induced tension, or by a combination of both. However, some evidence appears to suggest that this may be a misconception and that a solar sail could conceivably be heavily wrinkled and remain suitably reflective. Only in-flight data will provide conclusive resolution. It is immediately clear that this distinction is of critical importance, because if a sail film can be wrinkled then the deployment structure need only support and not tension the film, whereas if the film cannot be wrinkled then a much heavier structure is required to apply a tension. Once deployed the sail film must be oriented to direct the solar radiation pressure force for orbit maneuvering. Due to the large moment of inertia of a solar sail innovative engineering is again required.

The sail film reflective layer is supported by a substrate. The substrate is required principally to allow handling, folding, packing and deployment. The substrate must be coated with a suitable reflecting material for efficient photon reflection, and typically aluminum is favored. A further front coating, such as silicon oxide, may also be required in order to reduce pre-launch oxidation of the reflecting surface with a resultant loss of reflectivity. Alternatively, an ultraviolet-induced sublimation layer could be added to prevent pre-launch oxidation without adding mass to the actual solar sail flight mass. The sail substrate must have sufficient strength not to fail and create tears that may propagate during deployment or when fully deployed and under tension, if tension is required following deployment. Furthermore, since the reflective coating on the sail film will not have perfect reflectivity, a fraction of the incident solar radiation will be absorbed by the substrate and this must be dissipated through a thermally emitting rear surface coating. The choice of a suitable, high emissivity rear-surface coating is yet another design decision. Thin-film chromium, with an emissivity of the order of 0.64 appears to be a suitable candidate.

The sail substrate contributes a significant proportion of the total sail assembly mass, particularly for a large sail where the substrate mass dominates the sail's parasitic mass breakdown. The production of very thin films with good mechanical and thermal properties is thus central to solar sail realization. There is extensive industrial experience of the manufacture, coating and handling of thin films for a

number of ground and space applications. For example, primary spacecraft insulation is typically provided by multi-layer insulation (MLI) blankets which are constructed of alternative layers of aluminum coated Mylar[®] or Kapton[®] and a thin net of material such as nylon, Dacron[®], Nomex[®] or bridal veil. Note, however, that currently the typical thickness of commercially available thin films is excessive for moderate performance solar sails. Mylar is commercially available down to a thickness of only 0.9 μm , but it has a low resistance to solar ultraviolet radiation and so is unsuitable for long-duration exposure without double-sided coatings. Several thin film materials have been considered as potential sail substrates. Until recently the optimal sail film was generally considered to be Kapton. This does not have a melting point as such, but it does suffer a phase transition (glass transition temperature) above approximately 680 K. The safe, long-term maximum operating temperature for solar sail applications is thus generally considered to be between 520–570 K. It is this thermal limit which gives rise to the widely accepted minimum heliocentric radius of 0.25 au for solar sailing, although of course this does not take into account the thermal limits of the sail booms and other structural components, nor does it account for the thermal limits of the attached spacecraft. An all-aluminum sail film (that is to say one with no substrate) actually has a very similar minimum heliocentric distance even though bulk aluminum has a much higher melting point. The production of sail film of the order of 2 μm has recently been identified as a key technology requirement for mid-term solar sailing [10]. Such thin films are not routinely used for large volume commercial purposes, mainly due to the difficulty of handling them during manufacture. In addition to solar sails, other space applications such as solar concentrators and space telescope sunshades also require films thinner than commercially available Kapton. To this end, NASA and SRS Technologies have produced Clear Plastic-1 (CP-1) film down to a gauge of order 2 μm . This has very similar properties to Kapton film, making it highly suitable for solar sail applications. Indeed, CP-1 film is now generally accepted to be the sail film of choice.

A significant amount of development has been conducted since the 1990s into solar sail technology, and the deployment of large gossamer structures. Much of this technology development has focused on the sail film supporting booms that must be deployed post-launch, and has spawned a range of solutions from CFRP bistable booms to inflatable booms. Figure 11.23 shows a solar sail demonstrator made by L'Garde that was deployed in a large thermal vacuum chamber under ambient space conditions at NASA's Glenn Research Center in 2005. This same technology was selected for flight in 2011 as a NASA technology demonstration mission. The solar sail of the L'Garde Technology Demonstration Mission, named Sunjammer, will have seven

times the area (1,200 m^2) of the first sail flown in space, JAXA's IKAROS sail (Interplanetary Kite-craft Accelerated by Radiation Of the Sun), and at 32 kg it will weigh 10 times less. The L'Garde solar sail will produce a maximum thrust of approximately 0.01 N, giving it a sail loading (see later) of just over 25 g/m^2 and a characteristic acceleration (see later) of approximately 0.25 mm/s^2 . The Sunjammer sail is shown in Fig. 11.24 undergoing a single quadrant deployment test.

11.9.2.1 A Quantum Description of Solar Radiation Pressure

Using quantum mechanics, radiation pressure can be visualized as momentum transported by photons impacting on and then reflecting off a surface. As previously stated in Chap. 4, the term 'photon' was coined by Gilbert N. Lewis in a letter to *Nature* magazine in 1926 [11, 12]. From Planck's law, a photon of frequency ν will transport the energy given by

$$E = h\nu \quad (11.29)$$

where h is Planck's constant. Using Special Relativity the total energy of a moving body may be written as

$$E^2 = m_0^2 c^4 + p^2 c^2 \quad (11.30)$$

where c is the speed of light. Since a photon has zero rest mass, its energy may be written as

$$E = pc. \quad (11.31)$$

Using the photon energy defined by Eqs. 11.29 and 11.31, the momentum transported by a single photon is

$$p = \frac{h\nu}{c}. \quad (11.32)$$

The pressure on a body is found through consideration of the momentum transported by a flux of photons. At distance r from the Sun the energy flux may be written in terms of the solar luminosity, L_S , and scaled by the Sun–Earth distance, giving

$$W = \frac{L_S}{4\pi R_e^2} \left(\frac{R_e}{r}\right)^2 = W_e \left(\frac{R_e}{r}\right)^2 \quad (11.33)$$

where R_e is the mean distance of the Earth from the Sun, i.e. equivalent to 1 au. From Eq. 11.33, the energy ΔE transported across a surface of area A , normal to the incident radiation, in time Δt is given by,

$$\Delta E = WA\Delta t. \quad (11.34)$$

From Eq. 11.31, the energy then transports momentum Δp

Fig. 11.23 The 10-m solar sail deployment test performed by L'Garde in the 30-m vacuum chamber at NASA's Glenn Research Center, Plum Brook Station in 2005. *Image NASA*

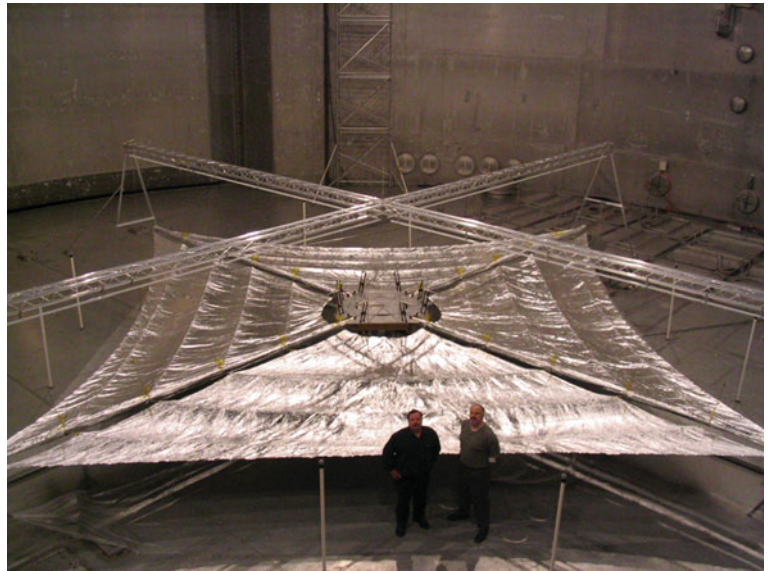


Fig. 11.24 A Sunjammer mission single-quadrant sail deployment test. *Image L'Garde, Inc*



$$\Delta p = \frac{\Delta E}{c}. \quad (11.35)$$

The pressure on the surface is thus defined as the momentum transported per unit time, per unit area, such that

$$P = \frac{1}{A} \left(\frac{\Delta p}{\Delta t} \right). \quad (11.36)$$

Accordingly, using Eq. 11.34 the pressure exerted on the surface due to momentum transport by photons is

$$P = \frac{W}{c}. \quad (11.37)$$

For a perfectly reflecting surface the actual pressure is twice the value given by Eq. 11.37, because momentum is transferred not only by incident photons but also by reflected photons, following Newton's second law. However, inefficiencies means that the factor will always be less than two.

11.9.2.2 Force on a Perfectly Reflecting Solar Sail

The acceleration experienced by a solar sail is a function of the attitude of the sail's reflective surface with respect to the Sun. For a solar sail, the force exerted on the surface due to incident photons is given by

$$\mathbf{F}_i = PA(\mathbf{u}_i \cdot \mathbf{n})\mathbf{u}_i. \quad (11.38)$$

The reflected photons will exert a force of equal magnitude on the surface, but in the specular reflected direction, $-\mathbf{u}_r$

$$\mathbf{F}_r = -PA(\mathbf{u}_i \cdot \mathbf{n})\mathbf{u}_r. \quad (11.39)$$

Noting that $\mathbf{u}_i - \mathbf{u}_r = 2(\mathbf{u}_i \cdot \mathbf{n})^2 \mathbf{n}$, the total force exerted on the solar sail is given by

$$\mathbf{F} = 2PA(\mathbf{u}_i \cdot \mathbf{n})^2 \mathbf{n}. \quad (11.40)$$

The total force may this be written as

$$\mathbf{F} = \frac{2AW_e}{c} \left(\frac{R_e}{r} \right)^2 (\mathbf{u}_i \cdot \mathbf{n})^2 \mathbf{n}. \quad (11.41)$$

The solar sail's performance is quantified by the total spacecraft mass per unit area (m/A) and is called the 'sail loading'. This is an important solar sail design parameter. The sail pitch angle is defined as the angle between the sail normal and the incident radiation. Using these definitions the solar sail acceleration may now be written as

$$\mathbf{a}_s = \frac{2W_e}{c} \frac{1}{\sigma} \left(\frac{R_e}{r} \right)^2 \cos^2(\alpha) \mathbf{n}. \quad (11.42)$$

The characteristic acceleration is defined as the actual acceleration experienced by the sail at a heliocentric distance of 1 au with the orientated sail normal to the Sun, such that $\alpha = 0$. The characteristic acceleration is a further design parameter and may be written as

$$\mathbf{a}_s = \frac{9.12\eta}{\sigma [g \text{ m}^{-2}]} [\text{mm s}^{-2}] \quad (11.43)$$

where an overall efficiency factor, η , is used to account for the finite reflectivity of the sail film. Typically the total solar sail efficiency is of the order of 85–90 %. It is important to note that although the efficiency of a sail does not alter its characteristic acceleration, it will alter the physical dimensions of the sail. The solar sail acceleration may also be written in terms of the solar gravitational acceleration as

$$\mathbf{a}_s = \beta \frac{GM_s}{r^2} (\hat{\mathbf{r}} \cdot \mathbf{n})^2 \mathbf{n}. \quad (11.44)$$

The dimensionless sail parameter β is defined as the ratio of the solar radiation pressure acceleration to the solar gravitational acceleration. This is called the 'lightness number' of the sail. The solar radiation pressure acceleration and the solar gravitational acceleration are both assumed to have an inverse square variation, therefore the lightness number is independent of the heliocentric distance of the sail. A more rigorous examination of the effect of radiation pressure on a surface can be found through the use of radiative transfer methods and an examination of the effect of a non-ideal reflecting solar sail, as in [13].

References

1. Millis, M.G., and Davis, E.W., (2009), "Frontiers of Propulsion Science", AIAA, Reston, VA.
2. Clark, J., (1972), "Ignition!", Rutgers University Press, New Brunswick, NJ.
3. Sutton, G.P., and Biblarz, O., "Rocket Propulsion Elements", Eight Edition, J.Wiley and Sons, New York, 2010.
4. Humble, R.W., Henry, G.N., and Larson, W.J., eds., (1995), "Space propulsion Analysis and Design", McGraw-Hill, New York.
5. Fearn, D.G., (2008), "Application of Ion Thrusters to High-Thrust, High-Specific-Impulse Nuclear Electric Missions", in: Nuclear Space Power and Propulsion Systems, ed. by C. Bruno, AIAA, Reston, VA.
6. Auweter-Kurtz, M., and Kurtz, H., (2008), "High-Power and High-Thrust-Density Electric Propulsion for In-Space Transportation", in: Nuclear Space Power and Propulsion Systems, by C. Bruno, ed., AIAA, Reston, VA, 2008.
7. Durante, M., and Bruno, C., (2010), "Impact of rocket propulsion technology on the radiation risk in missions to Mars", Eur. Phys. J. D, published online, 16 February 2010, DOI: [10.1140/epjd/e2010-00035-6](https://doi.org/10.1140/epjd/e2010-00035-6).
8. Turner, M.J.L., (2005), "Rocket and Space Propulsion", Springer-Praxis, Chichester.

9. Powell, J., Maise, G., and Paniagua, J., (2003), "HIP: a Hybrid NTP/NEP Propulsion System for Ultra Fast Robotic/Lander Missions to the Outer Solar System", IAC Paper presented at the 54th IAC, Bremen, Sept. 29 to Oct. 3, 2003.
10. Macdonald, Malcolm and McInnes, Colin (2011) "Solar sail science mission applications and advancement : solar sailing: concepts, technology, missions", *Advances in Space Research*, 48 (11). pp. 1702-1716. ISSN 0273-1177.
11. Lewis, G.N., Letter to the editor of Nature magazine, Vol. 118, Part 2, pp. 874-875, December 1926.
12. Griffiths, D. J. "The Photon (1900-1924)." Section 1.2 "Introduction to Elementary Particles". New York: Wiley, pp. 14-17, 1987.
13. McInnes, C.R., "Solar Sailing: Technology, Dynamics and Mission Applications", Springer-Praxis, Chichester, 1999.
24. Hill, P., and Peterson, C., (1992), "Mechanics and Thermodynamics of Propulsion", Addison- Wesley, Reading, MA.
25. Humble, R.W., Henry, G.N., and Larson, W.J., eds.,(1995), "Space propulsion Analysis and Design", McGraw-Hill, New York.
26. Huzel, D.K., and Huang, D.H., (1992), "Modern Engineering for Design of Liquid-Propellant Rocket Engines", AIAA, Reston, VA.
27. Jahn, R. G., Physics of electric propulsion, 1st ed., New York, McGraw-Hill, 1968.
28. Jahn, R. G. and Choueiri, E. Y., (2002), "Electric propulsion," in *Encyclopedia of Physical Science and Technology*, Vol. 5, 3rd ed., Academic Press, New York, pp. 125-141.
29. Koroteyev, A.S., Akimov, V.N., and Gafarov, A.A., (2007), "Creation and Perspectives of Application of Space Nuclear energy in Russia", *Polyot (Flight)*, Vol. 7, pp. 3-15.
30. Larson, W.J., Everett, D.F., and Puschell, J.J., (2011), "Space Mission Engineering The New SMAD", Microcosm Press.
31. Macdonald, Malcolm and McInnes, Colin (2011) "Solar sail science mission applications and advancement : solar sailing: concepts, technology, missions", *Advances in Space Research*, 48 (11). pp. 1702-1716. ISSN 0273-1177.

Further Reading

14. Brown, C.D., (1996), "Spacecraft Propulsion", AIAA, Reston, VA.
15. Bruno, C., and Accettura, A., eds., (2008), "Advanced Propulsion Systems and Technologies, Today to 2020", AIAA, Reston, VA.
16. Choueiri, E.Y., (2004), "A Critical History of Electric Propulsion: The First 50 Years (1906-1956)", *J. of Propulsion and Power*, Vol. 20, No. 2, March-April 2004, pp. 193-203.
17. Clark, J., (1972), "Ignition!", Rutgers University Press, New Brunswick, NJ.
18. Czysz, P.A., and Bruno, C., (2009), "Future Spacecraft Propulsion Systems", Springer-Praxis, London.
19. Durante, M., and Bruno, C., (2010), "Impact of rocket propulsion technology on the radiation risk in missions to Mars", *Eur. Phys. J. D*, published online, 16 February 2010, DOI: [10.1140/epjd/e2010-00035-6](https://doi.org/10.1140/epjd/e2010-00035-6).
20. Encrenaz, T., Bibring, J.-P., Blanc, M., Barucci, M.A., Roque, F., and Zarka, P., (2004), "The Solar System", Springer-Verlag, Berlin.
21. Friedman, L.,(1988),"Solar Sails and Interstellar Travel", J. Wiley and Sons, NY.
22. Goebel, D.M., and Katz, I., (2008), "Fundamentals of Electric Propulsion: Ion and Hall Thrusters", Jet Propulsion Laboratory, California Institute of Technology.
23. Gunn, S.V., and Ehresman, C.M., (2003), "The Space Propulsion Technology Base Established Four Decades Ago for the Thermal Nuclear Rocket is Ready for Current Applications", AIAA Paper 2003-4590, presented at the 39th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 20-23 July 2003, Huntsville, AL.
32. McInnes, C.R., "Solar Sailing: Technology, Dynamics and Mission Applications", Springer-Praxis, Chichester, 1999.
33. Powell, J., Maise, G., and Paniagua, J., (2003), "HIP: a Hybrid NTP/NEP Propulsion System for Ultra Fast Robotic/Lander Missions to the Outer Solar System", IAC Paper presented at the 54th IAC, Bremen, Sept. 29 to Oct. 3, 2003.
34. Stuhlinger, E., (1964), "Ion Propulsion for Space Flight", McGraw-Hill, New York.
35. Sutton, G.P., and Biblarz, O., "Rocket Propulsion Elements", Eight Edition, J.Wiley and Sons, New York, 2010.
36. Timnat, Y.M., (1987), "Advanced Chemical Rocket Propulsion", Academic Press, London.
37. Turner, M.J.L., (2005), "Rocket and Space Propulsion", Springer-Praxis, Chichester.
38. Wright, J.L., (1992), "Space Sailing", Gordon and Breach, Philadelphia.
39. Yang, V., Habiballah, M., Hulka, J. and Popp, M., eds., (2004), "Liquid Rocket Thrust Chambers: Aspects of Modeling, Analysis and Design", AIAA, Reston, VA.
40. Yang, V., and Anderson, W., eds., (1995), "Liquid Rocket Engine Combustion Instability", AIAA, Reston, VA.

This chapter describes space technology concepts and hardware associated with the spacecraft attitude and orbit control systems (AOCS). Practical examples of such systems as well as the fundamentals of AOCS analysis and design are emphasized throughout this chapter.

A typical AOCS architecture is illustrated in Fig. 12.1. The attitude determination and control system (ADCS) is one of the key subsystems of most spacecraft, and it consists of an attitude determination system (ADS) and an attitude control system (ACS). The ADCS provides the stabilization and control of the attitude (orientation) of a spacecraft using a variety of sensors and actuators in the presence of disturbance torques. The guidance, navigation, and control (GNC) system of a spacecraft is concerned with the orbital motion (trajectory) of the spacecraft's center of mass. The orbital GNC system is an area of space technology that plays a key role in the success of space missions that involve rendezvous, docking, and proximity operations. An integrated ADCS/GNC system is often referred to as the spacecraft AOCS. The GNC system and ADCS are analyzed and designed separately for most satellites, however. In order to accomplish various GNC tasks, the GNC system includes the sensors that provide measurements, the GNC software implemented in the on-board computer, and the actuators. The GNC software includes the navigation filter, the guidance algorithm, and the control algorithm, as illustrated in Fig. 12.1.

An in-depth and comprehensive treatment of spacecraft attitude determination theory and applications is provided in

[1]. Detailed descriptions of spacecraft guidance and control problems can be found in [2, 3]. Modern treatment of spacecraft attitude dynamics and control problems of practical interest can be found in [4, 5]. The ADCS/GNC systems are also discussed in [6–8] from the viewpoint of spacecraft systems engineering.

The fundamentals of attitude determination and control will be discussed in Sect. 12.1, and the ADCS will be further discussed in Sect. 12.2. The principles of orbital GNC systems will be treated in Sect. 12.3 with a detailed discussion of the orbital rendezvous problem.

12.1 Fundamentals of Attitude Determination and Control

12.1.1 Rotational Kinematics

The spacecraft attitude determination and control problem involves rotational kinematics. In this section, the rotational kinematics of a rigid body are considered to describe the orientation of a spacecraft that is in rotational motion. Throughout this section, the orientation of a reference frame fixed in a body is used to describe the orientation of the spacecraft body itself. This section is based on [5, Chap. 5].

12.1.1.1 Direction Cosine Matrix

Consider a reference frame A with a right-handed set of three orthogonal unit vectors $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ and a reference frame B with another right-handed set of three orthogonal unit vectors $\{\vec{b}_1, \vec{b}_2, \vec{b}_3\}$. The basis vectors $\{\vec{b}_1, \vec{b}_2, \vec{b}_3\}$ of B are expressed in terms of the basis vectors $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ of A as follows

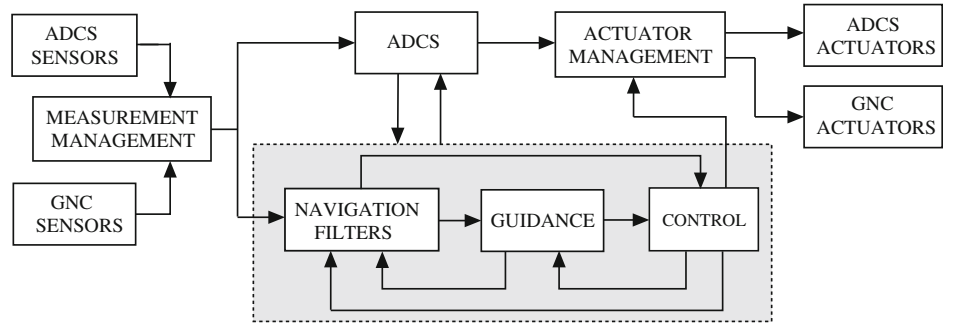
$$\begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vec{b}_3 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vec{a}_3 \end{bmatrix} = \mathbf{C}^{B/A} \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vec{a}_3 \end{bmatrix} \quad (12.1)$$

B. Wie (✉)
Iowa State University, Ames, Iowa, USA
e-mail: bongwie@iastate.edu

V. Lappas
University of Surrey Guildford, UK

J. Gil-Fernández
GMV, Tres Cantos, Spain

Fig. 12.1 Block diagram illustration of the spacecraft AOCs



where $\mathbf{C}^{B/A} \equiv [C_{ij}]$ is called the *direction cosine matrix* which describes the orientation of B relative to A . The direction cosine matrix $\mathbf{C}^{B/A}$ is also called the *rotation matrix* or *coordinate transformation matrix* to B from A . Such a coordinate transformation is symbolically represented as

$$\mathbf{C}^{B/A} : B \leftarrow A.$$

For brevity, \mathbf{C} for $\mathbf{C}^{B/A}$ is often used. Since each set of basis vectors of A and B consists of orthogonal unit vectors, the direction cosine matrix \mathbf{C} is an *orthonormal matrix*; thus

$$\mathbf{C}^{-1} = \mathbf{C}^T \quad (12.2)$$

which is equivalent to

$$\mathbf{C}\mathbf{C}^T = \mathbf{I} = \mathbf{C}^T\mathbf{C}. \quad (12.3)$$

In general, a square matrix \mathbf{A} is called an *orthogonal matrix* if $\mathbf{A}\mathbf{A}^T$ is a diagonal matrix, and it is called an *orthonormal matrix* if $\mathbf{A}\mathbf{A}^T$ is an identity matrix. For an orthonormal matrix \mathbf{A} , we have $\mathbf{A}^{-1} = \mathbf{A}^T$ and $|\mathbf{A}| = \pm 1$.

For an arbitrary vector \vec{r} expressed as

$$\vec{r} = y_1\vec{b}_1 + y_2\vec{b}_2 + y_3\vec{b}_3 = x_1\vec{a}_1 + x_2\vec{a}_2 + x_3\vec{a}_3 \quad (12.4)$$

the coordinate transformation relationship may be represented as

$$\mathbf{y} = \mathbf{C}\mathbf{x} \quad (12.5)$$

where \mathbf{C} is the direction cosine matrix of B relative to A , and \mathbf{y} and \mathbf{x} are the two corresponding component vectors defined as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Three elementary rotations about the 1st, 2nd, and 3rd axes, respectively, of the reference frame A are described by the following rotation matrices

$$\mathbf{C}_1(\theta_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_1 & \sin \theta_1 \\ 0 & -\sin \theta_1 & \cos \theta_1 \end{bmatrix} \quad (12.6a)$$

$$\mathbf{C}_2(\theta_2) = \begin{bmatrix} \cos \theta_2 & 0 & -\sin \theta_2 \\ 0 & 1 & 0 \\ \sin \theta_2 & 0 & \cos \theta_2 \end{bmatrix} \quad (12.6b)$$

$$\mathbf{C}_3(\theta_3) = \begin{bmatrix} \cos \theta_3 & \sin \theta_3 & 0 \\ -\sin \theta_3 & \cos \theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (12.6c)$$

where $\mathbf{C}_i(\theta_i)$ denotes the direction cosine matrix \mathbf{C} of an elementary rotation about the i th axis of A with an angle θ_i .

12.1.1.2 Euler Angles

One scheme for orienting a rigid body to a desired attitude is called a *body-axis rotation*; it involves successively rotating three times about the axes of the rotated, body-fixed reference frame. The first rotation is about any axis. The second rotation is about either of the two axes not used for the first rotation. The third rotation is then about either of the two axes not used for the second rotation. There are 12 sets of Euler angles for such successive rotations about the axes fixed in the body.

Consider three successive body-axis rotations to describe the orientation of a reference frame B relative to a reference frame A . A particular sequence chosen here is symbolically represented as

$$\mathbf{C}_1(\theta_1) \leftarrow \mathbf{C}_2(\theta_2) \leftarrow \mathbf{C}_3(\theta_3)$$

where $\mathbf{C}_i(\theta_i)$ indicates a rotation about the i th axis of the body-fixed frame with an angle θ_i .

The rotation matrix to B from A , or the direction cosine matrix of B relative to A , is then defined as

$$\begin{aligned} \mathbf{C}^{B/A} &\equiv \mathbf{C}_1(\theta_1)\mathbf{C}_2(\theta_2)\mathbf{C}_3(\theta_3). \\ &= \begin{bmatrix} c_2c_3 & c_2s_3 & -s_2 \\ s_1s_2c_3 - c_1s_3 & s_1s_2s_3 + c_1c_3 & s_1c_2 \\ c_1s_2c_3 + s_1s_3 & c_1s_2s_3 - s_1c_3 & c_1c_2 \end{bmatrix} \quad (12.7) \end{aligned}$$

where $c_i \equiv \cos \theta_i$ and $s_i \equiv \sin \theta_i$.

In general, there are 12 sets of Euler angles, each resulting in a different form for the rotation matrix $\mathbf{C}^{B/A}$. For example, the sequence of $\mathbf{C}_1(\theta_1) \leftarrow \mathbf{C}_3(\theta_3) \leftarrow \mathbf{C}_2(\theta_2)$ to B from A may be considered. For this case, the rotation matrix becomes

$$\begin{aligned} \mathbf{C}^{B/A} &\equiv \mathbf{C}_1(\theta_1)\mathbf{C}_3(\theta_3)\mathbf{C}_2(\theta_2) \\ &= \begin{bmatrix} c_2c_3 & s_3 & -s_2c_3 \\ -c_1c_2s_3 + s_1s_2 & c_1c_3 & c_1s_2s_3 + s_1c_2 \\ s_1c_2s_3 + c_1s_2 & -s_1c_3 & -s_1s_2s_3 + c_1c_2 \end{bmatrix}. \end{aligned} \quad (12.8)$$

In general, Euler angles have an advantage over direction cosines in that three Euler angles determine a unique orientation, although there is no unique set of Euler angles for a given orientation.

12.1.1.3 Quaternion

Consider Euler's eigenaxis rotation about an arbitrary axis fixed both in a body-fixed reference frame B and in an inertial reference frame A . A unit vector \vec{e} along the Euler axis is defined as

$$\begin{aligned} \vec{e} &= e_1\vec{a}_1 + e_2\vec{a}_2 + e_3\vec{a}_3 \\ &= e_1\vec{b}_1 + e_2\vec{b}_2 + e_3\vec{b}_3 \end{aligned}$$

where e_i are the direction cosines of the Euler axis relative to both A and B , and $e_1^2 + e_2^2 + e_3^2 = 1$.

Then the four Euler parameters or the quaternion can be defined as follows

$$q_1 = e_1 \sin(\theta/2) \quad (12.9a)$$

$$q_2 = e_2 \sin(\theta/2) \quad (12.9b)$$

$$q_3 = e_3 \sin(\theta/2) \quad (12.9c)$$

$$q_4 = \cos(\theta/2) \quad (12.9d)$$

where θ is the rotation angle about the Euler axis. Similar to the eigenaxis vector $\mathbf{e} = (e_1, e_2, e_3)$, a vector $\bar{\mathbf{q}} = (q_1, q_2, q_3)$ and the quaternion vector $\mathbf{q} = (q_1, q_2, q_3, q_4)$ can be defined such that

$$\bar{\mathbf{q}} = \mathbf{e} \sin \frac{\theta}{2} \quad (12.10)$$

$$\mathbf{q} = \begin{bmatrix} \bar{q} \\ q_4 \end{bmatrix}. \quad (12.11)$$

Note that the four Euler parameters are not independent of each other, but are constrained by the relationship

$$\mathbf{q}^T \mathbf{q} = \bar{\mathbf{q}}^T \bar{\mathbf{q}} + q_4^2 = q_1^2 + q_2^2 + q_3^2 + q_4^2 = 1 \quad (12.12)$$

The direction cosine matrix can also be parameterized in terms of the quaternion as follows

$$\begin{aligned} \mathbf{C}^{B/A} &= \mathbf{C}(\mathbf{q}) \\ &= \begin{bmatrix} 1 - 2(q_2^2 + q_3^2) & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_2q_1 - q_3q_4) & 1 - 2(q_1^2 + q_3^2) & 2(q_2q_3 + q_1q_4) \\ 2(q_3q_1 + q_2q_4) & 2(q_3q_2 - q_1q_4) & 1 - 2(q_1^2 + q_2^2) \end{bmatrix} \end{aligned} \quad (12.13)$$

which is often written as

$$\mathbf{C}(\mathbf{q}) = (q_4^2 - \bar{\mathbf{q}}^T \bar{\mathbf{q}})\mathbf{I} + 2\bar{\mathbf{q}}\bar{\mathbf{q}}^T - 2q_4\mathbf{Q} \quad (12.14)$$

where

$$\mathbf{Q} \equiv \begin{bmatrix} 0 & -q_3 & q_2 \\ q_3 & 0 & -q_1 \\ -q_2 & q_1 & 0 \end{bmatrix}. \quad (12.15)$$

Consider two successive rotations to A'' from A represented by

$$\mathbf{C}(\mathbf{q}') : A' \leftarrow A \quad (12.16a)$$

$$\mathbf{C}(\mathbf{q}'') : A'' \leftarrow A' \quad (12.16b)$$

where \mathbf{q}' is the quaternion associated with the coordinate transformation $A' \leftarrow A$ and \mathbf{q}'' is the quaternion associated with the coordinate transformation $A'' \leftarrow A'$. These successive rotations are also represented by a single rotation to A'' directly from A , as follows

$$\mathbf{C}(\mathbf{q}) : A'' \leftarrow A \quad (12.17)$$

where \mathbf{q} is the quaternion associated with the coordinate transformation $A'' \leftarrow A$, and

$$\mathbf{C}(\mathbf{q}) = \mathbf{C}(\mathbf{q}'')\mathbf{C}(\mathbf{q}'). \quad (12.18)$$

The resulting quaternion transformation relationship can be written as

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} q_4'' & q_3'' & -q_2'' & q_1'' \\ -q_3'' & q_4'' & q_1'' & q_2'' \\ q_2'' & -q_1'' & q_4'' & q_3'' \\ -q_1'' & -q_2'' & -q_3'' & q_4'' \end{bmatrix} \begin{bmatrix} q_1' \\ q_2' \\ q_3' \\ q_4' \end{bmatrix} \quad (12.19)$$

which is known as the quaternion multiplication rule in matrix form. The 4×4 orthonormal matrix in Eq. 12.19 is called the *quaternion matrix*. It can be written as

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} q_4' & -q_3' & q_2' & q_1' \\ q_3' & q_4' & -q_1' & q_2' \\ -q_2' & q_1' & q_4' & q_3' \\ -q_1' & -q_2' & -q_3' & q_4' \end{bmatrix} \begin{bmatrix} q_1'' \\ q_2'' \\ q_3'' \\ q_4'' \end{bmatrix}. \quad (12.20)$$

The 4×4 matrix in Eq. 12.20 is also orthonormal and is called the *quaternion transmuted matrix*.

12.1.1.4 Kinematic Differential Equations

Consider *kinematics* in which the relative orientation between two reference frames is time dependent. The time-dependent relationship between two reference frames is described by the so-called *kinematic differential equations*.

Consider two reference frames A and B , which are moving relative to each other. The angular velocity vector of a reference frame B with respect to a reference frame A is denoted by $\vec{\omega} \equiv \vec{\omega}^{B/A}$, and it is expressed in terms of the basis vectors of B as follows

$$\vec{\omega} = \omega_1 \vec{b}_1 + \omega_2 \vec{b}_2 + \omega_3 \vec{b}_3 = \begin{bmatrix} \vec{b}_1 & \vec{b}_2 & \vec{b}_3 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \quad (12.21)$$

where the angular velocity vector $\vec{\omega}$ is time dependent.

The kinematic differential equation for the direction cosine matrix \mathbf{C} is given by

$$\dot{\mathbf{C}} + \mathbf{\Omega} \mathbf{C} = 0 \quad (12.22)$$

where

$$\mathbf{\Omega} \equiv \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}. \quad (12.23)$$

Similar to the kinematic differential equation for the direction cosine matrix \mathbf{C} , the orientation of a reference frame B relative to a reference frame A can also be described by introducing the time dependence of Euler angles.

Consider the rotational sequence of $\mathbf{C}_1(\theta_1) \leftarrow \mathbf{C}_2(\theta_2) \leftarrow \mathbf{C}_3(\theta_3)$ to B from A . The time derivatives of Euler angles, called *Euler rates*, are denoted by $\dot{\theta}_3$, $\dot{\theta}_2$, and $\dot{\theta}_1$. These successive rotations result in

$$\begin{aligned} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} &= \begin{bmatrix} \dot{\theta}_1 \\ 0 \\ 0 \end{bmatrix} + \mathbf{C}_1(\theta_1) \begin{bmatrix} 0 \\ \dot{\theta}_2 \\ 0 \end{bmatrix} + \mathbf{C}_1(\theta_1) \mathbf{C}_2(\theta_2) \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & -\sin \theta_2 \\ 0 & \cos \theta_1 & \sin \theta_1 \cos \theta_2 \\ 0 & -\sin \theta_1 & \cos \theta_1 \cos \theta_2 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\theta}_3 \end{bmatrix}. \end{aligned} \quad (12.24)$$

Note that the 3×3 matrix in Eq. 12.24 is not an orthogonal matrix because \vec{b}_1 , \vec{a}_2'' , and \vec{a}_3' do not constitute a set of orthogonal unit vectors. The inverse relationship can be found by inverting the 3×3 nonorthogonal matrix in Eq. 12.24, as follows

$$\begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\theta}_3 \end{bmatrix} = \frac{1}{\cos \theta_2} \begin{bmatrix} \cos \theta_2 & \sin \theta_1 \sin \theta_2 & \cos \theta_1 \sin \theta_2 \\ 0 & \cos \theta_1 \cos \theta_2 & -\sin \theta_1 \cos \theta_2 \\ 0 & \sin \theta_1 & \cos \theta_1 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \quad (12.25)$$

which is the kinematic differential equation for the sequence of $\mathbf{C}_1(\theta_1) \leftarrow \mathbf{C}_2(\theta_2) \leftarrow \mathbf{C}_3(\theta_3)$.

If ω_1 , ω_2 , and ω_3 are known as functions of time, then the orientation of B relative to A as a function of time can be determined by solving Eq. 12.25. Numerical integration of Eq. 12.25, however, involves the computation of trigonometric functions of the angles. Also note that Eq. 12.25 becomes singular when $\theta_2 = \pi/2$. Such a mathematical singularity problem for a certain orientation angle can be avoided by selecting a different set of Euler angles, but it is an inherent property of all different sets of Euler angles.

The kinematic differential equation for the quaternion are given by

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \\ \dot{q}_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} q_4 & -q_3 & q_2 & q_1 \\ q_3 & q_4 & -q_1 & q_2 \\ -q_2 & q_1 & q_4 & q_3 \\ -q_1 & -q_2 & -q_3 & q_4 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ 0 \end{bmatrix} \quad (12.26)$$

which can be rewritten as

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \\ \dot{q}_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & \omega_3 & -\omega_2 & \omega_1 \\ -\omega_3 & 0 & \omega_1 & \omega_2 \\ \omega_2 & -\omega_1 & 0 & \omega_3 \\ -\omega_1 & -\omega_2 & -\omega_3 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix}. \quad (12.27)$$

In terms of $\bar{\mathbf{q}}$ and $\boldsymbol{\omega}$ defined as

$$\bar{\mathbf{q}} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}; \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}$$

the kinematic differential equation 12.27 can be rewritten as follows

$$\dot{\bar{\mathbf{q}}} = \frac{1}{2} (\boldsymbol{\omega} \times \bar{\mathbf{q}} - \bar{\mathbf{q}} \times \boldsymbol{\omega}) \quad (12.28a)$$

$$\dot{q}_4 = -\frac{1}{2} \boldsymbol{\omega}^T \bar{\mathbf{q}} \quad (12.28b)$$

where

$$\boldsymbol{\omega} \times \bar{\mathbf{q}} \equiv \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}.$$

In *strapdown inertial reference systems* of aerospace vehicles, the body rates, ω_1 , ω_2 , and ω_3 are measured by

rate gyros which are ‘strapped down’ to the vehicles. The kinematic differential equation 12.27 is then integrated numerically using an on-board flight computer to determine the orientation of the vehicles in terms of the quaternion. Inertial sensors such as star trackers or Sun sensors are employed to correct state propagation errors caused by angular-rate measurement uncertainties (e.g., gyro drift and bias).

The quaternion has no inherent geometrical singularity, unlike Euler angles. Moreover, the quaternion is well suited to on-board real-time computation because only products and no trigonometric relations exist in the quaternion kinematic differential equations. Thus, spacecraft orientation is now commonly described in terms of the quaternion.

12.1.2 Euler’s Rotational Equations of Motion

Consider a rigid spacecraft with a body-fixed reference frame B that has its origin at the center of mass. The angular velocity vector of the reference frame B with respect to an inertial reference frame A is denoted by $\vec{\omega} \equiv \vec{\omega}^{B/A}$, and it is expressed in terms of the basis vectors of B as follows

$$\vec{\omega} = \omega_1 \vec{b}_1 + \omega_2 \vec{b}_2 + \omega_3 \vec{b}_3 = \begin{bmatrix} \vec{b}_1 & \vec{b}_2 & \vec{b}_3 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}. \quad (12.29)$$

The angular momentum equation of a rigid body about its center of mass is

$$\vec{M} = \dot{\vec{H}} \quad (12.30)$$

where \vec{H} is the angular momentum vector of a rigid body about its mass center and \vec{M} is the external moment acting on the body about its mass center, expressed in terms of body-fixed basis vectors $\{\vec{b}_1, \vec{b}_2, \vec{b}_3\}$, as follows

$$\begin{aligned} \vec{H} &= H_1 \vec{b}_1 + H_2 \vec{b}_2 + H_3 \vec{b}_3 \\ \vec{M} &= M_1 \vec{b}_1 + M_2 \vec{b}_2 + M_3 \vec{b}_3. \end{aligned}$$

Furthermore

$$\dot{\vec{H}} \equiv \left\{ \frac{d\vec{H}}{dt} \right\}_A = \left\{ \frac{d\vec{H}}{dt} \right\}_B + \vec{\omega} \times \vec{H} \quad (12.31)$$

where

$$\left\{ \frac{d\vec{H}}{dt} \right\}_B = \dot{H}_1 \vec{b}_1 + \dot{H}_2 \vec{b}_2 + \dot{H}_3 \vec{b}_3. \quad (12.32)$$

The angular momentum vector is described by $\vec{H} = \hat{J} \cdot \vec{\omega}$ where \hat{J} is the inertia dyadic related to the inertia matrix as

$$\hat{J} = \begin{bmatrix} \vec{b}_1 & \vec{b}_2 & \vec{b}_3 \end{bmatrix} \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vec{b}_3 \end{bmatrix}. \quad (12.33)$$

The rotational equation of motion of a rigid body about its center of mass is then written as

$$\begin{aligned} \vec{M} &= \left\{ \frac{d\vec{H}}{dt} \right\}_B + \vec{\omega} \times \vec{H} \\ &= \left\{ \frac{d}{dt} (\hat{J} \cdot \vec{\omega}) \right\}_B + \vec{\omega} \times \hat{J} \cdot \vec{\omega} \\ &= \left\{ \frac{d\hat{J}}{dt} \right\}_B \cdot \vec{\omega} + \hat{J} \cdot \left\{ \frac{d\vec{\omega}}{dt} \right\}_B + \vec{\omega} \times \hat{J} \cdot \vec{\omega} \end{aligned} \quad (12.34)$$

where $\{d\hat{J}/dt\}_B = 0$ and $\{d\vec{\omega}/dt\}_B = \{d\vec{\omega}/dt\}_A = \dot{\vec{\omega}}$.

Finally

$$\vec{M} = \hat{J} \cdot \dot{\vec{\omega}} + \vec{\omega} \times \hat{J} \cdot \vec{\omega} \quad (12.35)$$

is called Euler’s rotational equation of motion in vector/dyadic form.

The rotational equation of motion in matrix form can also be obtained as follows

$$\begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} \dot{H}_1 \\ \dot{H}_2 \\ \dot{H}_3 \end{bmatrix} + \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}. \quad (12.36)$$

Since

$$\begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{33} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}$$

it is evident that

$$\begin{aligned} \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} &= \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \\ \dot{\omega}_3 \end{bmatrix} \\ &+ \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}. \end{aligned} \quad (12.37)$$

Defining a skew-symmetric matrix

$$\mathbf{\Omega} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} \quad (12.38)$$

it is seen that Eq. 12.37 can be rewritten concisely as

$$\mathbf{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\Omega}\mathbf{J}\boldsymbol{\omega} = \mathbf{M} \quad (12.39)$$

where

$$\mathbf{J} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix}; \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}; \quad \mathbf{M} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix}.$$

Using cross product notation of two column vectors, $\boldsymbol{\omega}$ and $\mathbf{J}\boldsymbol{\omega}$, defined as

$$\boldsymbol{\omega} \times \mathbf{J}\boldsymbol{\omega} \equiv \boldsymbol{\Omega}\mathbf{J}\boldsymbol{\omega}$$

it is possible to rewrite Eq. 12.39 as

$$\mathbf{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega} \times \mathbf{J}\boldsymbol{\omega} = \mathbf{M}. \quad (12.40)$$

For a principal-axis reference frame with a set of basis vectors $\{\vec{b}_1, \vec{b}_2, \vec{b}_3\}$, Euler's rotational equations of motion of a rigid body become

$$J_1\dot{\omega}_1 - (J_2 - J_3)\omega_2\omega_3 = M_1 \quad (12.41a)$$

$$J_2\dot{\omega}_2 - (J_3 - J_1)\omega_3\omega_1 = M_2 \quad (12.41b)$$

$$J_3\dot{\omega}_3 - (J_1 - J_2)\omega_1\omega_2 = M_3 \quad (12.41c)$$

where $J_1, J_2,$ and J_3 are the principal moments of inertia, $M_i = \vec{M} \cdot \vec{b}_i, \omega_i = \vec{\omega} \cdot \vec{b}_i$. These are three coupled, nonlinear ordinary differential equations for state variables $\omega_1, \omega_2,$ and ω_3 of a rigid body. These dynamical equations and the kinematic differential equations of the preceding sections completely describe the rotational motions of a rigid body with three rotational degrees of freedom (i.e., six state variables).

12.1.3 Attitude Determination Using Vector Observations

In this section an optimal attitude determination problem of finding the orthonormal matrix \mathbf{C} to minimize the least-squares loss function is considered

$$L = \frac{1}{2} \sum_{i=1}^n a_i |\mathbf{b}_i - \mathbf{C}\mathbf{r}_i|^2$$

where $\{\mathbf{r}_1, \dots, \mathbf{r}_n\}$ are a set of known reference unit vectors (e.g., the direction of the Earth, the Sun, a star, or the geomagnetic field) in the inertial frame, $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ are a set of the corresponding measured (observed) unit vectors in the spacecraft body-fixed frame, and n is the total number of measurements. Because these vectors are inaccurate, the weighting coefficients a_i are to be properly chosen to find the least-squares estimate of \mathbf{C} . This problem was first

posed by Wahba in 1965 [9] and is often referred to in the literature as Wahba's problem.

12.1.3.1 TRIAD Algorithm

Consider an attitude determination problem using only two vector measurements. The problem here is to determine the direction cosine matrix \mathbf{C} for the two observation unit vectors, \mathbf{b}_1 and \mathbf{b}_2 , corresponding to the two nonparallel reference unit vectors, \mathbf{r}_1 and \mathbf{r}_2 , as follows

$$\mathbf{b}_1 \leftarrow \mathbf{C}\mathbf{r}_1; \quad \mathbf{b}_2 \leftarrow \mathbf{C}\mathbf{r}_2. \quad (12.42)$$

Assuming that $\{\mathbf{r}_1, \mathbf{b}_1\}$ are more accurate than $\{\mathbf{r}_2, \mathbf{b}_2\}$, two sets of new basis vectors can be defined as

$$\mathbf{x}_1 = \mathbf{r}_1; \quad \mathbf{x}_2 = \frac{\mathbf{r}_1 \times \mathbf{r}_2}{|\mathbf{r}_1 \times \mathbf{r}_2|}; \quad \mathbf{x}_3 = \mathbf{x}_1 \times \mathbf{x}_2 \quad (12.43)$$

$$\mathbf{y}_1 = \mathbf{b}_1; \quad \mathbf{y}_2 = \frac{\mathbf{b}_1 \times \mathbf{b}_2}{|\mathbf{b}_1 \times \mathbf{b}_2|}; \quad \mathbf{y}_3 = \mathbf{y}_1 \times \mathbf{y}_2. \quad (12.44)$$

These are two triads of orthonormal unit vectors. There exists a unique orthogonal matrix \mathbf{C} which satisfies

$$\mathbf{y}_i = \mathbf{C}\mathbf{x}_i; \quad i = 1, 2, 3 \quad (12.45)$$

or

$$[\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3] = \mathbf{C}[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]. \quad (12.46)$$

The solution of Eq. 12.46 is then obtained as

$$\mathbf{C} = [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3][\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]^T = \sum_{i=1}^3 \mathbf{y}_i \mathbf{x}_i^T \quad (12.47)$$

since $[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]$ is an orthonormal matrix.

The algebraic method for determining a direction cosine matrix using only two vector measurements as discussed above is the so-called TRIAD algorithm described in [1, pp. 424–425, 10–12]. It was first proposed in [10] for a simple, but non-optimal, estimation of \mathbf{C} using two vector observations. Note that $\mathbf{C}\mathbf{r}_1$ becomes exactly \mathbf{b}_1 and that $\mathbf{C}\mathbf{r}_2$ doesn't become exactly \mathbf{b}_2 . The necessary and sufficient condition for $\mathbf{b}_2 \equiv \mathbf{C}\mathbf{r}_2$ is

$$\mathbf{r}_1^T \mathbf{r}_2 = \mathbf{b}_1^T \mathbf{b}_2. \quad (12.48)$$

As discussed in [1, pp. 426–428], a major drawback of the TRIAD algorithm is its *ad hoc* nature. The two measurements are heuristically combined to obtain an attitude estimate but the combination is not optimal in any statistical sense. Although the TRIAD algorithm has been implemented in numerous space missions, it cannot be easily applied to star trackers that provide many simultaneous vector measurements. The error covariance matrix associated with an estimated \mathbf{C} is often computed in terms of Euler angles [12]. Efficient methods for computing the covariance matrix of the TRIAD algorithm are presented in [11, 12].

12.1.3.2 QUEST Algorithm

The direction cosine matrix can also be determined when many simultaneous vector measurements are available [12].

Consider an optimal attitude determination problem of finding the orthonormal matrix \mathbf{C} to minimize the least-squares loss function

$$L = \frac{1}{2} \sum_{i=1}^n a_i |\mathbf{b}_i - \mathbf{C}\mathbf{r}_i|^2 \quad (12.49)$$

where the non-negative weighting coefficients are normalized as

$$\sum_{i=1}^n a_i = 1. \quad (12.50)$$

The problem of minimizing L can be transformed to the problem of maximizing the gain function G defined as

$$G = 1 - L = \frac{1}{2} \sum_{i=1}^n a_i \mathbf{b}_i^T \mathbf{C}\mathbf{r}_i = \frac{1}{2} \sum_{i=1}^n a_i \text{tr}[\mathbf{b}_i^T \mathbf{C}\mathbf{r}_i] = \text{tr}[\mathbf{C}\mathbf{B}^T] \quad (12.51)$$

where “tr” denotes the trace operator and the attitude profile matrix \mathbf{B} is defined as

$$\mathbf{B} = \sum_{i=1}^n a_i \mathbf{b}_i \mathbf{r}_i^T. \quad (12.52)$$

Because the nine elements of \mathbf{C} are subject to six constraints, it is better to parameterize \mathbf{C} in terms of the quaternion, as described by Eq. 12.14. Using the quaternion vector defined as

$$\mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{q}} \\ q_4 \end{bmatrix}$$

the gain function G can be obtained as

$$G = \mathbf{q}^T \mathbf{K} \mathbf{q} \quad (12.53)$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{S} - \sigma \mathbf{I} & \mathbf{z} \\ \mathbf{z}^T & \sigma \end{bmatrix}$$

and

$$\sigma = \text{tr}[\mathbf{B}] = \sum_{i=1}^n a_i \mathbf{b}_i^T \mathbf{r}_i \quad (12.54)$$

$$\mathbf{S} = \mathbf{B} + \mathbf{B}^T = \sum_{i=1}^n a_i (\mathbf{b}_i \mathbf{r}_i^T + \mathbf{r}_i \mathbf{b}_i^T) \quad (12.55)$$

$$\begin{aligned} \mathbf{z} &= \sum_{i=1}^n a_i (\mathbf{b}_i \times \mathbf{r}_i) \\ &= [B_{23} - B_{32} \quad B_{31} - B_{13} \quad B_{12} - B_{21}]^T. \end{aligned} \quad (12.56)$$

The problem of determining \mathbf{C} is transformed to finding the quaternion vector \mathbf{q} that maximizes the gain function G expressed by Eq. 12.53, subject to

$$\mathbf{q}^T \mathbf{q} = q_1^2 + q_2^2 + q_3^2 + q_4^2 = 1. \quad (12.57)$$

Adjoining this constraint to the gain function G via a Lagrange multiplier λ gives

$$G = \mathbf{q}^T \mathbf{K} \mathbf{q} - \lambda (\mathbf{q}^T \mathbf{q} - 1). \quad (12.58)$$

By letting the first differential of this new gain function with respect to \mathbf{q} be zero, the necessary condition is obtained as

$$\mathbf{K} \mathbf{q} = \lambda \mathbf{q}. \quad (12.59)$$

Thus, the eigenvector of \mathbf{K} becomes the optimal estimation of \mathbf{q} . The eigenvector associated with the largest eigenvalue maximizes the gain function G because

$$G = \mathbf{q}^T \mathbf{K} \mathbf{q} = \mathbf{q}^T \lambda \mathbf{q} = \lambda \mathbf{q}^T \mathbf{q} = \lambda. \quad (12.60)$$

The elegant computational algorithm described above is referred to as Davenport’s \mathbf{q} -method in the literature [1, pp. 426–428, 11, 12]. It provides an optimal least-squares estimate of \mathbf{C} by solving the eigenvalue/eigenvector problem of the matrix \mathbf{K} to find the optimal quaternion.

For the purpose of computationally efficient on-board implementation of the \mathbf{q} -method, the QUEST (QUaternion ESTimator) algorithm was proposed in [12], which is briefly introduced in this section as follows. As the gain function and loss function are related as

$$G = \sum_{i=1}^n a_i - L = 1 - L \quad (12.61)$$

it follows that

$$\lambda_{max} = 1 - L. \quad (12.62)$$

Thus, a good approximation of the optimal eigenvalue is

$$\lambda_{max} \approx 1. \quad (12.63)$$

Once the optimal eigenvalue λ_{max} is found using a Newton–Raphson iteration starting with 1 as the initial estimate, the next step of the QUEST algorithm is to solve the following eigenvector problem

$$\mathbf{K} \mathbf{q} = \lambda_{max} \mathbf{q} \quad (12.64)$$

where \mathbf{q} is the optimal quaternion. For the QUEST algorithm, this eigenvector equation is rewritten as

$$\mathbf{p} = [(\lambda_{max} + \sigma)\mathbf{I} - \mathbf{S}]^{-1}\mathbf{z} \quad (12.65)$$

where \mathbf{p} is the Gibbs vector or the Rodriguez parameters defined as

$$\mathbf{p} = \frac{\bar{\mathbf{q}}}{q_4} = \mathbf{e} \tan(\theta/2). \quad (12.66)$$

Instead of inverting the matrix in Eq. 12.65, Gaussian elimination may be used to solve the following equation

$$[(\lambda_{max} + \sigma)\mathbf{I} - \mathbf{S}]\mathbf{p} = \mathbf{z}. \quad (12.67)$$

After finding the optimal \mathbf{p} , the optimal quaternion is then simply obtained as

$$\mathbf{q} = \frac{1}{\sqrt{1 + \mathbf{p}^T \mathbf{p}}} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}. \quad (12.68)$$

A method of avoiding the singularity when $\theta = \pi$ is also described in [12]. Further detailed discussions of the QUEST method and its 3×3 quaternion covariance matrix can be found in [12]. A different way of finding the direction cosine matrix that minimizes Wahba's loss function using the singular value decomposition of the matrix \mathbf{B} is presented in [13].

12.1.4 Recursive Attitude Determination

The attitude determination problem considered in the preceding section is a static, single-frame attitude determination problem in which all of the vector measurements are made at the same attitude. In this section, a general recursive estimation problem of time-varying attitude is considered. Although QUEST-based recursive methods have been developed in [14–16], Kalman filtering is the primary means of real-time spacecraft attitude estimation in the presence of various sensor errors [17–20].

12.1.4.1 QUEST-Based Recursive Methods

A simple solution to the time-varying attitude estimation problem was proposed in [14]. It is based on propagating and updating the matrix \mathbf{B} , as follows

$$\mathbf{B}(t_j) = \mu \Phi_{3 \times 3}(t_j, t_{j-1}) \mathbf{B}(t_{j-1}) + \sum_{i=1}^{n_j} a_i \mathbf{b}_i \mathbf{r}_i^T \quad (12.69)$$

where $\Phi_{3 \times 3}(t_j, t_{j-1})$ is the state transition matrix of the attitude rotation matrix \mathbf{C} , $\mu < 1$ is a fading memory factor to be properly chosen, and n_j is the total number of vector observations at time t_j . The optimal attitude estimate at t_j is then computed by the QUEST algorithm for $\mathbf{B}(t_j)$.

The REQUEST algorithm proposed in [15] propagates and updates Davenport's matrix \mathbf{K} as

$$\mathbf{K}(t_j) = \mu \Phi(t_j, t_{j-1}) \mathbf{K}(t_{j-1}) \Phi^T(t_j, t_{j-1}) + \sum_{i=1}^{n_j} a_i \mathbf{K}_i \quad (12.70)$$

where $\Phi(t_j, t_{j-1})$ is the 4×4 quaternion state transition matrix and \mathbf{K}_i is Davenport's matrix \mathbf{K} for a single observation.

A major disadvantage of such QUEST-based recursive methods is the use of a simple fading memory scalar approximation for the sensor and process noises. The performance of the QUEST-based methods has been significantly improved by an extended QUEST algorithm [15, 16]. In [16], the fading memory factor is optimized using a statistical cost function recasting the REQUEST algorithm in a statistical filtering framework. The Extended QUEST algorithm solves the attitude estimation problem by finding the attitude quaternion \mathbf{q}_j and the auxiliary state vector \mathbf{x}_j , which minimize the loss function

$$\begin{aligned} J(\mathbf{q}_j, \mathbf{x}_j) = & \frac{1}{2} \sum_{i=1}^{n_j} \sigma_i^{-2} |\mathbf{b}_i - \mathbf{C}(\mathbf{q}_j) \mathbf{r}_i|^2 + \frac{1}{2} |\mathbf{R}_{ww(j-1)} \mathbf{w}_{j-1}|^2 \\ & + \frac{1}{2} |\mathbf{R}_{qq(j-1)} (\mathbf{q}_{j-1} - \hat{\mathbf{q}}_{j-1})|^2 \\ & + \frac{1}{2} |\mathbf{R}_{xq(j-1)} (\mathbf{q}_{j-1} - \hat{\mathbf{q}}_{j-1}) + \mathbf{R}_{xx(j-1)} (\mathbf{x}_{j-1} - \hat{\mathbf{x}}_{j-1})|^2 \end{aligned} \quad (12.71)$$

subject to the quaternion propagation equation

$$\mathbf{q}_j = \Phi(t_j, t_{j-1}; \mathbf{q}_{j-1}, \mathbf{x}_{j-1}, \mathbf{w}_{j-1}) \mathbf{q}_{j-1} \quad (12.72)$$

the auxiliary state filter propagation equation

$$\mathbf{x}_j = \mathbf{f}_x(t_j, t_{j-1}; \mathbf{q}_{j-1}, \mathbf{x}_{j-1}, \mathbf{w}_{j-1}) \quad (12.73)$$

and the quaternion normalization constraint $|\mathbf{q}_j|^2 = \mathbf{q}_j^T \mathbf{q}_j = 1$. The a posteriori estimates of \mathbf{q} and \mathbf{x} at time t_{j-1} are denoted as \mathbf{q}_{j-1} and \mathbf{x}_{j-1} , respectively, the process noise vector is denoted by \mathbf{w}_{j-1} , and the standard deviations associated with \mathbf{b}_i measurements are denoted by σ_i .

The Extended QUEST algorithm employs two separate computational phases [15].

In the dynamic propagation phase

$$\tilde{\mathbf{q}}_j = \Phi(t_j, t_{j-1}; \hat{\mathbf{q}}_{j-1}, \hat{\mathbf{x}}_{j-1}, 0) \hat{\mathbf{q}}_{j-1} \quad (12.74)$$

$$\tilde{\mathbf{x}}_j = \mathbf{f}_x(t_j, t_{j-1}; \hat{\mathbf{q}}_{j-1}, \hat{\mathbf{x}}_{j-1}, 0). \quad (12.75)$$

The loss function after the propagation phase becomes

$$\begin{aligned} J(\mathbf{q}_j, \mathbf{x}_j) = & \frac{1}{2} \sum_{i=1}^{n_j} \sigma_i^{-2} |\mathbf{b}_i - \mathbf{C}(\mathbf{q}_j) \mathbf{r}_i|^2 + \frac{1}{2} |\tilde{\mathbf{R}}_{qq,j} (\mathbf{q}_j - \tilde{\mathbf{q}}_{j-1})|^2 \\ & + \frac{1}{2} |\tilde{\mathbf{R}}_{xq,j} (\mathbf{q}_j - \tilde{\mathbf{q}}_j) + \tilde{\mathbf{R}}_{xx,j} (\mathbf{x}_j - \tilde{\mathbf{x}}_j)|^2 \end{aligned} \quad (12.76)$$

where $\tilde{\mathbf{R}}$ matrices are computed via a QR factorization.

In the measurement update phase, the optimal \mathbf{x}_j is given by

$$\mathbf{x}_j = \tilde{\mathbf{x}}_j - \tilde{\mathbf{R}}_{\mathbf{xx},j}^{-1} \tilde{\mathbf{R}}_{\mathbf{xq},j} (\mathbf{q}_j - \tilde{\mathbf{q}}_j) \quad (12.77)$$

and the loss function, Eq. 12.76, becomes

$$J(\mathbf{q}_j, \hat{\mathbf{x}}_j) = -\mathbf{q}_j^T \left[\sum_{i=1}^{n_j} \sigma_i^{-2} \mathbf{K}_i \right] \mathbf{q}_j + \frac{1}{2} |\tilde{\mathbf{R}}_{\mathbf{q},j} (\mathbf{q}_j - \tilde{\mathbf{q}}_{j-1})|^2. \quad (12.78)$$

The best estimate $\hat{\mathbf{q}}_j$ is then obtained by minimizing this modified loss function, Eq. 12.78. The best estimate $\hat{\mathbf{x}}_j$ is then obtained as

$$\hat{\mathbf{x}}_j = \tilde{\mathbf{x}}_j - \tilde{\mathbf{R}}_{\mathbf{xx},j}^{-1} \tilde{\mathbf{R}}_{\mathbf{xq},j} (\hat{\mathbf{q}}_j - \tilde{\mathbf{q}}_j). \quad (12.79)$$

The details of the Extended QUEST can be found in [15, 16].

12.1.4.2 Extended Kalman Filtering

A variety of recursive attitude estimation algorithms based on Kalman filtering, extended Kalman filtering, unscented Kalman filtering, or particle filtering, can be found in [17–20]. The Kalman filter was originally developed in 1960 as a new approach to linear filtering and prediction problems. When it is applied to non-linear dynamical systems, it is then referred to as the extended Kalman filter (EKF). The principle of the EKF is briefly introduced here.

Consider a nonlinear dynamical system described by

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, t) + \mathbf{G}(t)\mathbf{w}(t) \quad (12.80)$$

where \mathbf{x} is the state vector and \mathbf{w} is the process noise vector. It is assumed that the process noise is a Gaussian white noise whose mean and covariance are characterized as

$$E[\mathbf{w}(t)] = 0 \quad (12.81)$$

$$E[\mathbf{w}(t)\mathbf{w}^T(\tau)] = \mathbf{Q}(t)\delta(t - \tau). \quad (12.82)$$

The initial mean values of the state vector and the initial covariance of the state estimation error vector are given by

$$E[\mathbf{x}(t_0)] = \hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0 \quad (12.83)$$

$$E\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}][\mathbf{x}(t_0) - \hat{\mathbf{x}}]^T\} = \mathbf{P}(t_0) = \mathbf{P}_0. \quad (12.84)$$

The estimated state vector satisfies the following equation

$$\dot{\hat{\mathbf{x}}} = E[\mathbf{f}(\mathbf{x}, t)] = \hat{\mathbf{f}}(\hat{\mathbf{x}}, t) \approx \mathbf{f}(\hat{\mathbf{x}}, t) \quad (12.85)$$

and its solution is expressed as

$$\hat{\mathbf{x}}(t) = \Phi(t, \hat{\mathbf{x}}(t_0), t_0). \quad (12.86)$$

Let the state estimation error vector and its error covariance matrix be defined as

$$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t). \quad (12.87)$$

$$\mathbf{P}(t) = E[\tilde{\mathbf{x}}(t)\tilde{\mathbf{x}}^T(t)] \quad (12.88)$$

Then

$$\dot{\tilde{\mathbf{x}}} \approx \mathbf{F}(t)\tilde{\mathbf{x}}(t) + \mathbf{G}(t)\mathbf{w}(t) \quad (12.89)$$

where

$$\mathbf{F}(t) = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}(t)}. \quad (12.90)$$

The solution of Eq. 12.89 is given by

$$\tilde{\mathbf{x}}(t) = \Phi(t, t_0)\tilde{\mathbf{x}}(t_0) + \int_{t_0}^t \Phi(t, \tau)\mathbf{G}(\tau)\mathbf{w}(\tau)d\tau \quad (12.91)$$

where $\Phi(t, t_0)$ is the state transition matrix with the following properties

$$\frac{\partial}{\partial t} \Phi(t, t_0) = \mathbf{F}(t)\Phi(t, t_0) \quad (12.92)$$

$$\Phi(t_0, t_0) = \mathbf{I}. \quad (12.93)$$

The error covariance matrix $\mathbf{P}(t)$ satisfies the Riccati equation

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^T(t). \quad (12.94)$$

The solution of the Riccati equation is given by

$$\mathbf{P}(t) = \Phi(t, t_0)\mathbf{P}(t_0)\Phi^T(t, t_0) + \int_{t_0}^t \Phi(t, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\Phi^T(t, \tau)d\tau. \quad (12.95)$$

The estimated state vector and the state estimation error covariance matrix are then propagated as

$$\hat{\mathbf{x}}_j^- = \Phi(t_j, \hat{\mathbf{x}}_{j-1}^+, t_{j-1}) \quad (12.96)$$

$$\mathbf{P}_j^- = \Phi(t_j, t_{j-1})\mathbf{P}_{j-1}^+\Phi^T(t_j, t_{j-1}) + \mathbf{N}_{j-1} \quad (12.97)$$

where

$$\mathbf{N}_{j-1} = \int_{t_{j-1}}^{t_j} \Phi(t_j, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\Phi^T(t_j, \tau)d\tau. \quad (12.98)$$

A measurement model is given by

$$\mathbf{y}_j = \mathbf{h}(\mathbf{x}_j) + \mathbf{v}_j \quad (12.99)$$

with

$$E[\mathbf{v}_j] = 0 \quad (12.100)$$

$$E[\mathbf{v}_j \mathbf{v}_j^T] = \mathbf{R}_j \quad (12.101)$$

and its measurement sensitivity matrix is obtained as

$$\mathbf{H}_j = \left. \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}_j}. \quad (12.102)$$

The minimum-variance estimate of \mathbf{x}_j using the measurement \mathbf{y}_j is updated as

$$\hat{\mathbf{x}}_j^+ = \hat{\mathbf{x}}_j^- + \mathbf{K}_j[\mathbf{y}_j - \mathbf{h}(\hat{\mathbf{x}}_j^-)] \quad (12.103)$$

where the Kalman filter gain matrix is given by

$$\mathbf{K}_j = \mathbf{P}_j^- \mathbf{H}_j^T [\mathbf{H}_j \mathbf{P}_j^- \mathbf{H}_j^T + \mathbf{R}_j]^{-1}. \quad (12.104)$$

The error covariance matrix is updated as

$$\mathbf{P}_j^+ = [\mathbf{I} - \mathbf{K}_j \mathbf{H}_j] \mathbf{P}_j^-. \quad (12.105)$$

The detailed applications of the EKF to the spacecraft attitude estimation problem with the state vector consisting of the attitude quaternion, the gyro bias vector, and other uncertain parameters can be found in [17–20].

12.1.4.3 Unscented Kalman Filtering

The EKF is widely employed for the state estimation of nonlinear dynamical systems. However, the unscented Kalman filtering (UKF) is known to perform better than the EKF because the UKF reduces the linearization errors of the EKF. The UKF algorithm [18–20] is briefly described as follows.

Consider a discrete-time nonlinear system described by

$$\mathbf{x}_{j+1} = \mathbf{f}(\mathbf{x}_j, j) + \mathbf{w}_j \quad (12.106a)$$

$$\mathbf{y}_j = \mathbf{h}(\mathbf{x}_j, j) + \mathbf{v}_j \quad (12.106b)$$

where \mathbf{x}_j is the state vector, \mathbf{y}_j is the measurement vector, \mathbf{w}_j is the process noise vector, and \mathbf{v}_j is the measurement noise vector. It is assumed that \mathbf{w}_j and \mathbf{v}_j are zero-mean uncorrelated Gaussian noise processes with covariance matrices \mathbf{Q}_j and \mathbf{R}_j , respectively.

The UKF is initialized as

$$\hat{\mathbf{x}}_0^+ = E[\mathbf{x}_0] \quad (12.107a)$$

$$\mathbf{P}_0^+ = E[(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+)(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+)^T]. \quad (12.107b)$$

The next step is to obtain a set of sigma points using the current best estimate of the mean and covariance as

$$\hat{\mathbf{x}}_{j-1}^i = \hat{\mathbf{x}}_{j-1}^+ + \tilde{\mathbf{x}}_{j-1}^i \quad (12.108a)$$

$$\tilde{\mathbf{x}}_{j-1}^i = \left[\sqrt{n \mathbf{P}_{j-1}^+} \right]_i^T; \quad i = 1, \dots, n \quad (12.108b)$$

$$\tilde{\mathbf{x}}_{j-1}^{n+i} = - \left[\sqrt{n \mathbf{P}_{j-1}^+} \right]_i^T; \quad i = 1, \dots, n. \quad (12.108c)$$

Using the propagated sigma point vectors $\hat{\mathbf{x}}_j^i$, we obtain a priori state estimate $\hat{\mathbf{x}}_j^-$ and error covariance \mathbf{P}_j^- as

$$\hat{\mathbf{x}}_j^i = \mathbf{f}(\hat{\mathbf{x}}_{j-1}^i, j) \quad (12.109a)$$

$$\hat{\mathbf{x}}_j^- = \frac{1}{2n} \sum_{i=1}^{2n} \hat{\mathbf{x}}_j^i a_i \quad (12.109b)$$

$$\mathbf{P}_j^- = \frac{1}{2n} \sum_{i=1}^{2n} (\hat{\mathbf{x}}_j^i - \hat{\mathbf{x}}_j^-)(\hat{\mathbf{x}}_j^i - \hat{\mathbf{x}}_j^-)^T + \mathbf{Q}_{j-1} \quad (12.109c)$$

where a_i are weighting coefficients. Sigma points are recomputed using the current best estimate of the mean and covariance, as follows

$$\hat{\mathbf{x}}_j^i = \hat{\mathbf{x}}_j^- + \tilde{\mathbf{x}}_j^i \quad (12.110a)$$

$$\tilde{\mathbf{x}}_j^i = \left[\sqrt{n \mathbf{P}_j^-} \right]_i^T; \quad i = 1, \dots, n \quad (12.110b)$$

$$\tilde{\mathbf{x}}_j^{n+i} = - \left[\sqrt{n \mathbf{P}_j^-} \right]_i^T; \quad i = 1, \dots, n. \quad (12.110c)$$

The predicted observation vector \mathbf{y}_j and the covariance matrices are computed as

$$\hat{\mathbf{y}}_j^i = \mathbf{h}(\hat{\mathbf{x}}_j^i, j) \quad (12.111a)$$

$$\hat{\mathbf{y}}_j = \frac{1}{2n} \sum_{i=1}^{2n} \hat{\mathbf{y}}_j^i \quad (12.111b)$$

$$\mathbf{P}_{y(j)} = \frac{1}{2n} \sum_{i=1}^{2n} (\hat{\mathbf{y}}_j^i - \hat{\mathbf{y}}_j)(\hat{\mathbf{y}}_j^i - \hat{\mathbf{y}}_j)^T + \mathbf{R}_j \quad (12.111c)$$

$$\mathbf{P}_{xy(j)} = \frac{1}{2n} \sum_{i=1}^{2n} (\hat{\mathbf{x}}_j^i - \hat{\mathbf{x}}_j^-)(\hat{\mathbf{y}}_j^i - \hat{\mathbf{y}}_j)^T. \quad (12.111d)$$

Similar to the Kalman filter, the a posteriori state vector $\hat{\mathbf{x}}_j^+$ is updated using the measurement vector \mathbf{y}_j

$$\hat{\mathbf{x}}_j^+ = \hat{\mathbf{x}}_j^- + \mathbf{K}_j(\mathbf{y}_j - \hat{\mathbf{y}}_j) \quad (12.112a)$$

$$\mathbf{K}_j = \mathbf{P}_{xy(j)} \mathbf{P}_{y(j)}^{-1} \quad (12.112b)$$

$$\mathbf{P}_j^+ = \mathbf{P}_j^- - \mathbf{K}_j \mathbf{P}_{y(j)} \mathbf{K}_j^T \quad (12.112c)$$

12.1.5 Introduction to Spacecraft Attitude Control

12.1.5.1 Quaternion Feedback Control

Most three-axis stabilized spacecraft use a sequence of rotational maneuvers about each control axis. Many spacecraft also perform rotational maneuvers about an inertially fixed axis during an acquisition mode (e.g., Sun-acquisition or Earth-acquisition) so that a particular sensor will pick up a particular target. Spacecraft are sometimes required to maneuver as fast as possible within the physical limits of their actuators and sensors. Because quaternions are well suited for on-board real-time computation, spacecraft orientation is nowadays commonly described in terms of the quaternions. A simple quaternion-feedback control logic for three-axis, large-angle reorientation maneuvers [3] is briefly introduced here.

Consider the attitude dynamics of a rigid spacecraft described by Euler's rotational equation of motion of the form

$$\mathbf{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega} \times \mathbf{J}\boldsymbol{\omega} = \mathbf{u} \quad (12.113)$$

where $\mathbf{u} = (u_1, u_2, u_3)$ is the control torque input vector. The quaternion kinematic differential equations are given by

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \\ \dot{q}_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & \omega_3 & -\omega_2 & \omega_1 \\ -\omega_3 & 0 & \omega_1 & \omega_2 \\ \omega_2 & -\omega_1 & 0 & \omega_3 \\ -\omega_1 & -\omega_2 & -\omega_3 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix}. \quad (12.114)$$

A linear state feedback controller of the following form can be employed for large-angle reorientation maneuvers

$$\mathbf{u} = -\mathbf{K}\bar{\mathbf{q}}_e - \mathbf{C}\boldsymbol{\omega} \quad (12.115)$$

where $\bar{\mathbf{q}}_e = (q_{1e}, q_{2e}, q_{3e})$ is the attitude-error quaternion vector and \mathbf{K} and \mathbf{C} are controller gain matrices to be determined. The controller gain matrices can be simply selected as $\mathbf{K} = k\mathbf{J}$ and $\mathbf{C} = c\mathbf{J}$ where k and c are positive scalars to be properly chosen [3]. The attitude-error quaternions $(q_{1e}, q_{2e}, q_{3e}, q_{4e})$ are computed using the desired or commanded attitude quaternions $(q_{1c}, q_{2c}, q_{3c}, q_{4c})$ and the current attitude quaternions (q_1, q_2, q_3, q_4) [3], as follows

$$\begin{bmatrix} q_{1e} \\ q_{2e} \\ q_{3e} \\ q_{4e} \end{bmatrix} = \begin{bmatrix} q_{4c} & q_{3c} & -q_{2c} & -q_{1c} \\ -q_{3c} & q_{4c} & q_{1c} & -q_{2c} \\ q_{2c} & -q_{1c} & q_{4c} & -q_{3c} \\ q_{1c} & q_{2c} & q_{3c} & q_{4c} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} \quad (12.116)$$

If the commanded attitude quaternion vector is simply the origin defined as

$$(q_{1c}, q_{2c}, q_{3c}, q_{4c}) = (0, 0, 0, +1)$$

then the control logic of Eq. 12.115 becomes

$$\mathbf{u} = -\mathbf{K}\bar{\mathbf{q}} - \mathbf{C}\boldsymbol{\omega} \quad (12.117)$$

where $\bar{\mathbf{q}} = (q_1, q_2, q_3)$. Detailed discussions and applications of the quaternion-feedback control logic can be found in [3].

12.1.5.2 PID: Classical Proportional-Integral-Derivative Control

The PID (proportional-integral-derivative) control logic is commonly used in most feedback controllers, including spacecraft attitude control systems. To illustrate the basic concept of the PID control, consider a typical single-axis attitude control problem of spacecraft. This so-called double integrator plant is described by

$$J\ddot{\theta}(t) = u(t) + w(t) \quad (12.118)$$

where J is the spacecraft inertia, θ is the spacecraft attitude, u is the control torque, and w is the external disturbance torque.

Assuming that the spacecraft attitude and angular rate can be directly measured, a standard PD (proportional-derivative) control logic can be expressed as

$$u(t) = -K_P\theta(t) - K_D\dot{\theta}(t) \quad (12.119)$$

where K_P and K_D are controller gains to be properly determined. The closed-loop system is then described by

$$J\ddot{\theta}(t) + K_D\dot{\theta}(t) + K_P\theta(t) = w(t). \quad (12.120)$$

The closed-loop characteristic equation of the system is

$$Js^2 + K_Ds + K_P = 0.$$

The control design task is to properly determine K_P and K_D to meet given performance/stability specifications of the closed-loop system. Let ω_n and ζ be the desired natural frequency and damping ratio of the closed-loop poles. Then the desired closed-loop characteristic equation becomes

$$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0$$

and the controller gains K_P and K_D can be determined as

$$K_P = J\omega_n^2 \quad (12.121a)$$

$$K_D = 2J\zeta\omega_n. \quad (12.121b)$$

The damping ratio ζ is often selected as $0.5 \leq \zeta \leq 0.707$, and the natural frequency ω_n is then considered as the *bandwidth* of the PD controller of a system with a rigid-body mode. For a unit-step disturbance, this closed-loop system with the PD controller results in a non-zero steady-state attitude $\theta(\infty) = 1/K_P$. However, the

steady-state attitude error $\theta(\infty)$ can be made small by designing a high-bandwidth control system.

In order to keep the desired attitude to be $\theta = 0$ at steady state in the presence of a constant disturbance, consider a PID controller of the form

$$u(t) = -K_P\theta(t) - K_I \int \theta(t)dt - K_D\dot{\theta}(t). \quad (12.122)$$

It can be shown that for a constant disturbance, the closed-loop system with the PID controller, in fact, results in a zero steady-state output $\theta(\infty) = 0$.

The closed-loop characteristic equation can be found as

$$Js^3 + K_Ds^2 + K_Ps + K_I = 0$$

and the desired closed-loop characteristic equation can be expressed as

$$(s^2 + 2\zeta\omega_n s + \omega_n^2)(s + 1/T) = 0$$

where ω_n and ζ denote, respectively, the natural frequency and damping ratio of the complex poles associated with the rigid-body mode, and T is the time constant of the real pole associated with integral control. The PID controller gains can then be determined as

$$K_P = J \left(\omega_n^2 + \frac{2\zeta\omega_n}{T} \right) \quad (12.123a)$$

$$K_I = J \frac{\omega_n^2}{T} \quad (12.123b)$$

$$K_D = J \left(2\zeta\omega_n + \frac{1}{T} \right). \quad (12.123c)$$

The time constant T of integral control is often selected as

$$T \approx \frac{10}{\zeta\omega_n}.$$

12.1.5.3 Modern State-Feedback Control Design

Consider a linearized spacecraft dynamics model as described by the following state-space equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (12.124)$$

where \mathbf{x} is the state vector and \mathbf{u} is the control input vector. The linear state-feedback control law is assumed as

$$\mathbf{u} = -\mathbf{K}\mathbf{x}. \quad (12.125)$$

The gain matrix \mathbf{K} of the state feedback control logic can be determined by minimizing the linear quadratic performance index

$$J = \frac{1}{2} \int_0^\infty (\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}) dt \quad (12.126)$$

where \mathbf{Q} is the state weighting matrix and \mathbf{R} is the control input weighting matrix. The gain matrix \mathbf{K} is then obtained as

$$\mathbf{K} = \mathbf{R}^{-1} \mathbf{B}^T \mathbf{X} \quad (12.127)$$

by solving the algebraic Riccati equation

$$0 = \mathbf{A}^T \mathbf{X} + \mathbf{X} \mathbf{A} - \mathbf{X} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{X} + \mathbf{Q}. \quad (12.128)$$

For further details of linear-quadratic regulator (LQR) control theory and applications, see [21].

If the actual state is not available for state-feedback control, a spacecraft's dynamics can be considered to be described by the following state-space equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{w} \quad (12.129a)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v} \quad (12.129b)$$

where \mathbf{w} is the process noise and \mathbf{v} is the measurement noise. Both \mathbf{w} and \mathbf{v} are assumed to be white noise processes with

$$E[\mathbf{w}(t)\mathbf{w}^T(\tau)] = \mathbf{W}\delta(t - \tau)$$

$$E[\mathbf{v}(t)\mathbf{v}^T(\tau)] = \mathbf{V}\delta(t - \tau)$$

where \mathbf{W} and \mathbf{V} are the corresponding spectral density matrices. The estimated-state feedback controller is then described by

$$\dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u} + \mathbf{L}(\mathbf{y} - \mathbf{C}\hat{\mathbf{x}}) \quad (12.130)$$

$$\mathbf{u} = -\mathbf{K}\hat{\mathbf{x}} \quad (12.131)$$

where $\hat{\mathbf{x}}$ is the estimated state for feedback control, \mathbf{K} is the regulator gain matrix, and \mathbf{L} the estimator gain matrix. The controller gain matrix \mathbf{K} is determined by Eq. 12.127, and the estimator gain matrix \mathbf{L} of the linear-quadratic-estimator (LQE) is selected such that the observation error

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}} \quad (12.132)$$

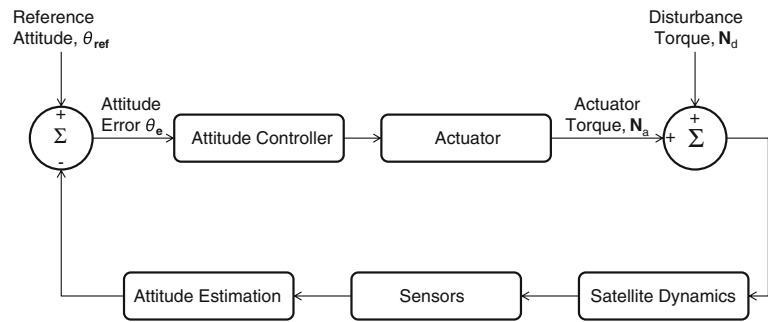
is minimized in the presence of noise, which is done by solving the algebraic Riccati equation

$$0 = \mathbf{A}\mathbf{Y} + \mathbf{Y}\mathbf{A}^T - \mathbf{Y}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\mathbf{Y} + \mathbf{G}\mathbf{W}\mathbf{G}^T \quad (12.133)$$

where \mathbf{Y} is the estimate-error covariance matrix, and then \mathbf{L} is computed as

$$\mathbf{L} = \mathbf{Y}\mathbf{C}^T\mathbf{V}^{-1}. \quad (12.134)$$

Fig. 12.2 Attitude control system, control diagram. *Image* Malcolm Macdonald



A more detailed treatment of linear-quadratic-gaussian (LQG) control theory and applications to aerospace dynamical systems can be found in [21, 22].

12.2 ADCS: Attitude Determination and Control Systems

The attitude determination and control system (ADCS) stabilizes, controls and positions a satellite in a desired orientation despite any external or internal disturbances acting on it. The satellite's payload requires a specific pointing direction whether the payload is a camera, a science instrument, or an antenna. Satellites also require orientation for thermal control, or to acquire the Sun for their solar panels. The ADCS system uses sensors in order to determine a satellite's attitude and actuators to control the vehicle to a required direction. The ADCS also needs to achieve the various mission and payload objectives such as pointing accuracy, stability, rotational rate (slew) and sensing with many physical constraints such as mass, power, volume, computer power/storage, space environment, robustness/lifetime and cost. As previously stated, the ADCS is a synthesis of two subsystems the attitude determination system (ADS) and the attitude control system (ACS) that controls the attitude/motion of a satellite as depicted in Fig. 12.2.

Space mission requirements, satellite size, cost and the space environment lead to different and diverse choices for the selection of ADCS hardware and control schemes. This section provides a top-level insight into the design of a practical ADCS. Additional information on ADCS technology can be found in Refs. [1, 4–8].

12.2.1 Requirements and Stabilization Methods

Satellites come with very diverse attitude determination and control requirements depending on their mission, orbit, and

payload. For example, Earth observation satellites require very high levels of pointing and stability in order to ensure that their images are not blurry, or in order to efficiently transmit on-board data to specific ground stations. Actuators such as reaction wheels, thrusters or electro-magnets (magnetorquers) which can be used to maneuver a spacecraft (change its attitude) can cause disturbance torques themselves that require careful selection and design of both actuators and sensors. Table 12.1 lists the typical ADCS requirements for a satellite mission.

Having established the ADCS requirements, it is important to select the way that a spacecraft will be controlled. There are several methods of controlling a spacecraft.

12.2.1.1 Gravity Gradient

Gravity gradient stabilization exploits Newton's law of general gravitation, and through the use of gravitational forces can always keep a spacecraft nadir-pointing. This is achieved by using a boom to extend a small distinct mass (usually a magnetometer in order to minimize magnetic interference) from the spacecraft, which becomes the second distinct mass, by a distance of 3–6 m. These two masses, which are connected by a thin and light boom, can then be used to exploit the difference in gravitational pull on the main satellite platform and the additional mass (magnetometer, say) due to the difference in their distance from Earth. This small difference can be sufficient to enable the satellite/additional mass system to be aligned with the radius vector at all times as an orbiting pendulum. The gravity gradient stabilization scheme can be beneficial for coarse pointing ($\sim 5^\circ$) around the nadir axis, while the other two axes will still need to be stabilized. This method of stabilization was first exploited by the US Department of Defense Gravity Experiment (DODGE) satellite in 1967, which captured the first color full-Earth image; see Fig. 12.3. Gravity gradient stabilization was also used on early UoSAT satellites, from Surrey Satellite Technology Ltd. (SSTL) in the UK, in the 1980s, which were used to store and forward communications [23].

Table 12.1 ADCS performance requirements [8, 38]

Requirement	Definition	Example
<i>Determination</i>		
Accuracy/attitude knowledge	How well a satellite's orientation is with respect to an absolute reference	0.1°, 3- σ
Range	Range of angular motion over which accuracy must be met	Attitude attained within 30° of nadir
<i>Control</i>		
Accuracy	How well the satellite attitude can be controlled with respect to a commanded direction	0.1°, 3- σ ; includes determination and control errors
Range	Range of angular motion over which control performance must be met	Full range, within 30° of nadir, 20° of sun
Stability/jitter	A specified angle bound or angular rate limit on short-term, high frequency motion	0.1°/s or a required value to keep spacecraft motion from blurring sensor/imager data
Slew rate/agility	Slew or angular rate required to perform a rapid maneuver	3°/s
Drift	A limit on slow, low frequency vehicle motion	1°/hr
Settling time	Allowed time to recover from maneuvers or upsets	2° maximum rotation, used to limit nutation, wobbling

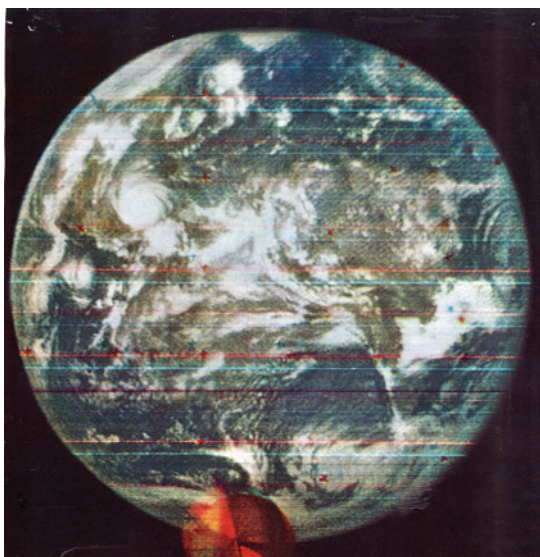


Fig. 12.3 The first full face color portrait of the earth taken by the DODGE satellite in 1967; taken at an altitude of 29,000 km. The hurricane above the gulf of Mexico is the Beulah hurricane. The image was taken with a black and white TV camera, which took three photos with a red, green and a blue filter to create the color image. The small disc in front of the picture is a colour match card. *Image US DoD*

12.2.1.2 Magnetic Stabilization

By approximating the Earth's magnetic field as a dipole, it is possible to have a satellite use a magnetometer to track the Earth's magnetic field lines in a 'compass mode' thus allows the vehicle to be passively stabilized but with coarse attitude (5–10°) due to the various irregularities and harmonics of the Earth's magnetic field [1, 4].

12.2.1.3 Spin Stabilization

Spinning a satellite generates an angular momentum vector that remains nearly fixed in inertial space. The angular

momentum generated provides gyroscopic stiffness to a spinning satellite, making it less prone to external disturbances and more stable for thruster/apogee motor firings. This passive stabilization technique was popular in the 1970s with GEO communication satellites and is still used during the cruise phase of interplanetary missions due to its simplicity and systems benefits for thermal and communication purposes. Detailed dynamical formulations for spinners can be found in Refs. [1, 4–8] (Fig. 12.4).

12.2.1.4 Dual Spin Stabilization

As a variation of the spin based stabilization scheme, a dual spin satellite has two parts of its structure spinning at different angular rates about the same axis. In this case, one section of the satellite spins to provide angular momentum, while the other part (platform) is de-spun and points in a fixed direction, for example towards the Earth. Such a scheme can be beneficial for a spacecraft in which the structure (diameter) of the platform is required to be 'thin' to fit in a launch vehicle fairing. The disadvantage of this scheme is the added complexity for carrying bearings and slip rings between the rotating parts of the satellite. Dual spin satellites were also popular in the 1970s, in particular for GEO satellites where the high gain antennas could stay fixed towards Earth.

12.2.1.5 Bias Momentum Stabilization

As a 3-axis system, a momentum bias system uses one actuator-momentum wheel aligned about the pitch axis normal to the orbit plane. Gyroscopic stiffness is used in order to control the vehicle by keeping the momentum wheel spinning at a constant rate. Small variations in wheel speed facilitate control of the pitch axis. Yaw-roll coupling for nadir-pointing bias momentum systems can be used to control the other two axes.



Fig. 12.4 The spin-stabilized Meteosat satellite integration at Aérospatiales Cannes facilities

12.2.1.6 Zero Momentum Stabilization

Most spacecraft today use the 3-axis zero momentum stabilization scheme because it can provide higher accuracy. In these systems, reaction wheels are used for each axis in order to compensate for external disturbances and to complete various commanded maneuvers. A pointing error is used to make the reaction wheels accelerate from an initial zero value and then the wheels move to a small spin rate that keeps increasing due to the maneuvers required and due to secular disturbances, taking them to their saturation limits. This requires a desaturation strategy, known as ‘momentum dumping’ or unloading, that uses magnetorquers or thrusters to enable the wheels to be spun-down.

12.2.2 An Attitude Determination System Example

As an example of a spacecraft attitude determination system (ADS), consider the Inertial Stellar Compass (ISC) developed by the Draper Laboratory for NASA’s New Millennium Program ST-6 project [24]. Its performance was successfully flight validated aboard the TacSat-2 satellite launched on December 16, 2006. The ISC is a miniature, low-power ADS developed for use with low-cost small satellites. It was designed to be suitable for a wide range of future missions because of its low mass, low power, and low volume design, and its self-initializing, autonomous operational capability. The ISC is composed of a wide field-of-view active-pixel star camera and micro-gyros, with associated data processing and power electronics, as illustrated in Fig. 12.5. Periodic updates from the star camera are used to correct the effect of gyro drift and bias on obtaining the attitude quaternion information from the rate gyros. The unique feature of the ISC is that those two miniaturized devices are integrated into a very low mass and low power

unit along with a microprocessor. It has a total mass of 2.5 kg, a power requirement of 3.5 W, and an accuracy of 0.1° (1 %).

12.3 Disturbance Torques

In order to design and size an ADCS, the torques acting on the spacecraft must be quantified in a similar fashion to the orbit perturbations discussed in Chap. 4. These can be divided into controlled actuator torques (e.g. magnetorquers, reaction wheels, control moment gyros, etc.) and external torques (e.g. gravity gradient, aerodynamic, solar pressure, etc.).

12.3.1 Gravity Gradient

The gravity gradient torque is a torque that originates from the ‘dumb bell’ effect on a long thin rotating object [1]. This torque derives from the finite distance between the opposite ends of the spacecraft, with a slight difference in the forces acting on those ends, resulting in a torque about the spacecraft’s center of mass. The gravity gradient torque for a satellites is defined [1, 4] as

$$\mathbf{N}_{\text{GG}} = \frac{3\mu}{R_c^5} \mathbf{R}_c \times (\mathbf{J} \cdot \mathbf{R}_c) \quad (12.135)$$

where, μ is Earth’s gravitational parameter, which was defined in Chap. 4 and has the value of approximately $3.986 \times 10^{14} \text{ m}^3 \text{ s}^{-2}$, \mathbf{R}_c is the position vector of the spacecraft’s mass center from the Earth’s center, R_c is the magnitude of \mathbf{R}_c , and \mathbf{J} is the spacecraft inertia matrix (or dyadic).

12.3.2 Solar Radiation Pressure Torque

This torque is caused mainly by the difference in location of the satellite’s center of pressure and its center of mass. Solar radiation will reflect off the satellite in parts of the spacecraft’s orbit and this will create a torque about the spacecraft’s center of mass. This torque is defined [1] as

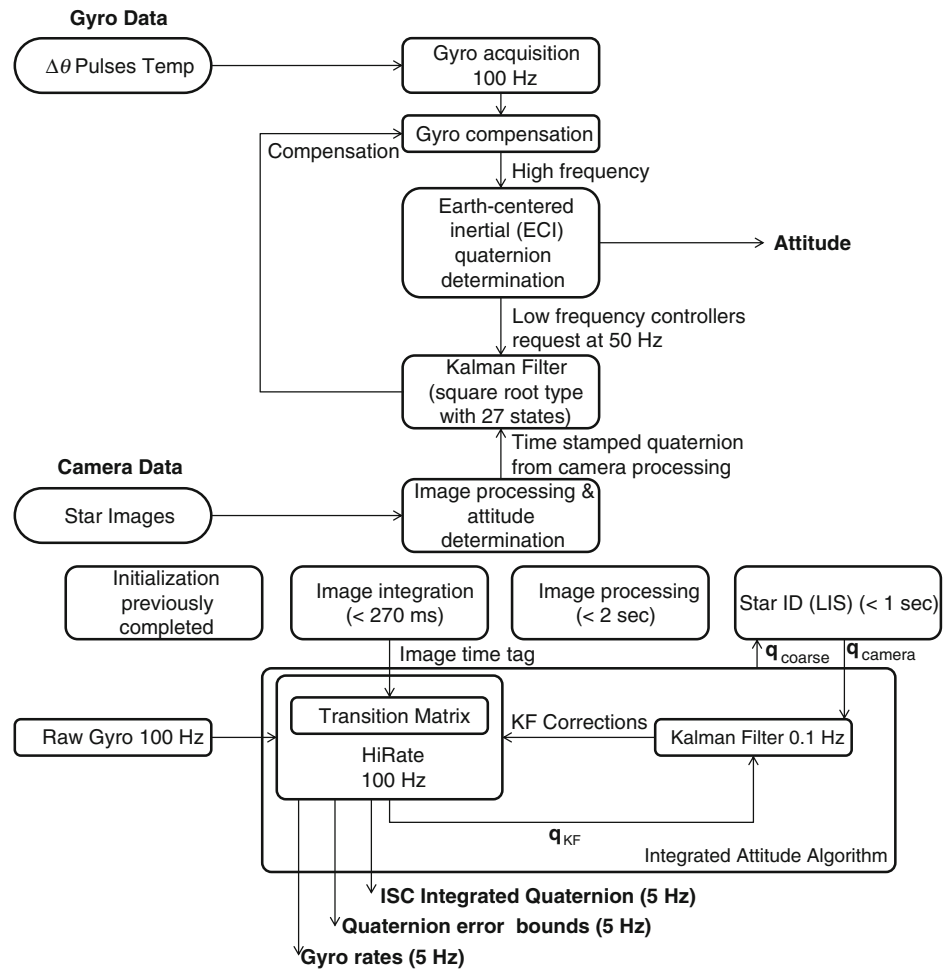
$$\mathbf{N}_{\text{sp}} = F(\mathbf{C}_{\text{ps}} - \mathbf{C}_{\text{g}}) \quad (12.136)$$

where

$$F = \frac{F_s}{c} A_s (1 + q) \cos(i), \quad (12.137)$$

F_s is the average solar irradiance (approximately $1,366 \text{ Wm}^{-2}$, the integrated power from ASTM E490—00a(2006) and ISO-21348 as discussed in Chap. 3), c is the

Fig. 12.5 Block diagram illustration of an attitude determination system, called the Inertial Stellar Compass (ISC), developed by Draper Laboratory under NASA's New Millennium ST-6 project [24]



speed of light (approximately $3.0 \times 10^8 \text{ m s}^{-1}$), \mathbf{C}_{ps} is the center of solar pressure vector, \mathbf{C}_g is the center of mass, A_s is the spacecraft surface area projected towards the Sun, i is the Sun incidence angle, and q is the reflectivity/transparency factor. For the $\mathbf{C}_{ps}-\mathbf{C}_g$ term, an estimated value of 0.1 m is used, and for the reflectivity q a value of 0.6 is typical for a small spacecraft [1].

12.3.3 Aerodynamic Disturbance Torque

In low earth orbits (LEO, i.e. <2,000 km), the effect of Earth's atmosphere (drag) must also be considered on the satellite's attitude. From Eq. 12.137, the atmospheric torque disturbance \mathbf{N}_A is directly proportional to the cross sectional area A_p and to atmospheric density ρ , i.e.

$$\mathbf{N}_A = \frac{1}{2} (\rho C_D A_p V^2) (\mathbf{C}_{pa} - \mathbf{C}_g) \quad (12.138)$$

where ρ is the atmospheric density (kg m^{-3}), C_D is the drag coefficient, A_p is the spacecraft projected area (m^2), V is the spacecraft velocity (m s^{-1}), \mathbf{C}_{pa} is the center of aerodynamic pressure of the spacecraft, and \mathbf{C}_g is the center of mass.

12.3.4 Internal Disturbance Torques

In addition to external disturbances, satellites can encounter internal disturbances. These can be generated by various factors that can, to some extent, be controlled through careful design

- Thruster misalignments, thruster output mismatch
- Moving components such as data recorders, pumps, stepper motors/mechanisms
- Liquid sloshing from propulsion tanks
- Thermal gradients/abrupt changes due to eclipses
- Dynamics, oscillatory resonances due to complex satellite structures and/or flexible appendages
- Interaction of current in spacecraft harnesses with external magnetic fields to create a magnetic torque.

12.4 Attitude Sensors

A suite of sensors is required to determine the attitude of a spacecraft, including its rates and angular position, despite constraints such as eclipses. The attitude information needs to be provided continuously with sufficient accuracy. There

are two categories of sensors: reference sensors provide a reference or a ‘datum’ of the direction of an object such as the Sun, a planet or a star, even though this could be interrupted by an eclipse; and inertial sensors provide continuous attitude readings, but due to errors are required to have attitude or calibration corrections from reference sensors so that the attitude error is kept within an acceptable tolerance. Due to the various sensor concepts and constraints, a combination reference and inertial sensors is used for a spacecraft using a balance of performance, mass/power consumption and cost.

12.4.1 Sun Sensors

Detecting the presence and/or the orientation of the Sun relative to the spacecraft is important in most space missions. Be it to time the thrust pulses for attitude control, or to use the Sun as one of the reference directions for determining the spacecraft’s orientation, or simply to maintain any sensitive components on-board in the shadow, Sun sensors have become routine. The Sun is a luminous body and can be approximated as a point source because at Earth’s heliocentric distance its arc radius is 0.267° . Therefore, it is relatively simple to detect and discriminate the Sun from other stars and planets. As a result, many Sun sensors have been manufactured over the years ranging from basic techniques that simply identify the presence/absence of the Sun, to sophisticated technologies that pinpoint the Sun’s direction to the accuracy of several hundredths of a degree. These sensors can be classified into three basic categories: the Sun presence detector, the analog Sun sensor, and the digital Sun sensor. The following sections briefly explain the operation of these sensors, including the hardware involved.

12.4.1.1 Sun Presence Detector

As the name indicates, the Sun presence detector outputs a signal when the Sun vector is in its field of observation. The configuration of the slits and the field of view of a sensor varies for the particular application for which the sensor is being used. Normally these sensors are used in cases when a particular component on-board is sensitive to Sun and has to be switched ON/OFF relative to the Sun’s presence or absence. For example, equipment such as space telescopes and star trackers need to be protected from direct sunlight. Various configurations of Sun-presence sensors have been developed [1].

12.4.1.2 Shadow Bar Sun Presence Detector

The shadow bar sensor detects the Sun when it is in the narrow field of view zone denoted α in Fig. 12.6, thereby

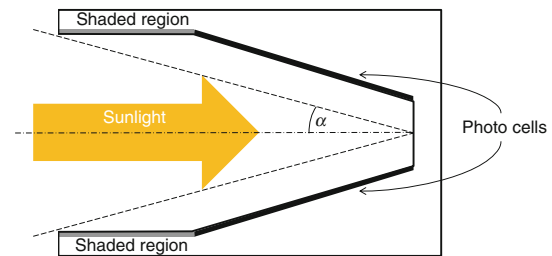


Fig. 12.6 Shadow bar Sun presence detector. *Image* Malcolm Macdonald

outputting a signal to the control devices which protect any co-located sensitive equipment.

12.4.1.3 Slit Sun Presence Detector

Here the photocell lies beneath a slit(s) and generates an output signal when the Sun vector lies on the plane of the slit(s). Two slit detectors are normally used in spin-stabilized platforms in order to detect the spin axis of the spacecraft. The principle of operation is simple, with a pulse output being generated when the following condition is satisfied.

$$\mathbf{n} \cdot \mathbf{s} = 0 \quad (12.139)$$

where \mathbf{n} is the vector normal to the slit plane and \mathbf{s} is the Sun vector in the frame of the sensor box. As the spacecraft spins, the normal vector rotates and Eq. 12.138 traces a cosine waveform.

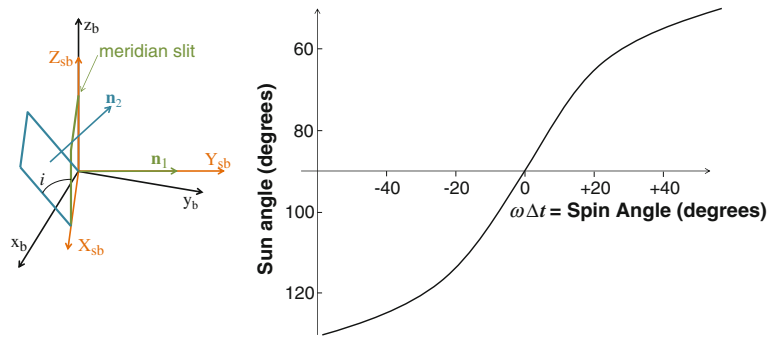
12.4.1.4 V-Slit Sun Presence Detector

A particular configuration of the Sun sensor called the V-slit sensor has been widely used to find the spin axis orientation of spin-stabilized spacecraft. A V-slit Sun sensor has two slits, the meridian slit parallel to the spin axis and the skew slit inclined at an angle i to the spin axis. During each rotation of the spacecraft, the Sun vector crosses the plane of the meridian slit once and the skew slit once. Denoting the Sun sensor reference frame as $[\mathbf{X}_{sb} \ \mathbf{Y}_{sb} \ \mathbf{Z}_{sb}]$ and the spacecraft body frame as $[\mathbf{x}_b \ \mathbf{y}_b \ \mathbf{z}_b]$ with \mathbf{z}_b as the spin-axis, then for the Sun vector to lie within the plane of each of the two slits, the dot-product of the slit’s normal vector and the Sun vector must be zero. Hence, the following condition has to be satisfied

$$\mathbf{n}_1 \cdot \mathbf{s} = 0 \quad \mathbf{n}_2 \cdot \mathbf{s} = 0 \quad (12.140)$$

where \mathbf{n}_1 and \mathbf{n}_2 are the vector normal to the slit planes and \mathbf{s} is the Sun vector in the frame of the sensor box. Figure 12.7 illustrates the geometry of the two V slits while the plot shows the angle between the spin axis and the Sun vector as a function of the spin angle with $i = 45^\circ$.

Fig. 12.7 V-slit Sun presence detector geometry (left) and Sun angle as a function of spin angle (right). Image Malcolm Macdonald



12.4.1.5 Single and Two Axis Analog Cosine Sensor

The basic principle of operation for an analog Sun sensor is that the total energy flux on the surface of the photo cell will be proportional to the cosine of the incidence angle of the Sun vector. As the current generated from the photo cell follows the cosine law, it is also known as the cosine law detector. Using one photo cell it is possible to compute a two dimensional representation of the Sun vector, thereby making it a one axis Sun sensor. A two axis Sun sensor is a combination of two single axis Sun sensors and it gives a complete three-dimensional representation for the Sun vector.

The basic design for a single axis Sun sensor requires only a single solar cell. To improve the performance, accuracy and the linear range of operation, many configurations of solar cells have been developed, although the basic principle remains the same. The following mathematical representation of the single and two axes Sun sensor gives a simple method for computation of the Sun vectors' orientation on-board the spacecraft

$$I(\alpha) = A \cos(\alpha) \quad (12.141)$$

where A is a constant that depends on the physical properties of the solar cell being used. Therefore, measuring the output current generated and the knowing the properties of the solar cell being used the incident angle α can be computed; however, the inverse cosine quadrant $-90^\circ \leq \alpha \leq 0^\circ$, or $0^\circ \geq \alpha \geq 90^\circ$ cannot be resolved. In order to choose the correct solution, another solar cell placed with its optical axis perpendicular to the first cell to provide another angle β

$$I(\beta) = A \cos(\beta). \quad (12.142)$$

As discussed in Chap. 10, the cosine law holds well for Sun angles ranging from 0° to about 50° , beyond which the electrical output deviates significantly from the cosine value. The actual power versus angle curve is called the Kelly cosine. The accuracy for the sensor will be less when the angle of incidence is around $\pm 90^\circ$, i.e. when the Sun vector is almost parallel to the photo cell, thereby restricting the operational range. In this case, a lookup table is used to

detect the angle of incidence, based on ground experimental data, in order to determine the Kelly cosine (Fig. 12.8).

12.4.1.6 Digital Sun Sensor

The digital Sun sensors provides higher accuracy than analog Sun sensors. In order to obtain a complete description of Sun vector orientation, two single axis digital Sun sensors with their optical planes at 90° are required. The digital Sun sensor comprises of an optical head and a signal-processing unit. The optical head has a narrow slit for sunlight to pass through and illuminate reticle slits that are organized to represent a suitable code such as a gray code or a binary code. A greater understanding in the choice and usage of gray binary code is given in [1, 4]. The encoded output from the reticle slits is decoded in order to obtain the Sun vector orientation using the signal-processing unit. An Automatic Threshold Adjust (ATA), half the width of the other photo cells is used. Therefore, the current generated would be half of any other illuminated slit. The current generated by the ATA is taken as a reference, to indicate whether any of the other slits are illuminated by the Sun. If the output voltage from any other slit is greater than twice the current generated by the ATA, then the corresponding bit is ON, otherwise it is OFF. The Sun vector incident angles α and β are computed after the coded bits are converted in the sensor electronics board.

With the evolution of Micro Electro-Mechanical System (MEMS) technology, miniature sensors using Active Pixel Sensors (APS) emerged. Photo cell detectors used in these devices can be based on CMOS and CCD technologies. A CMOS sensor module is comparatively smaller and has less power consumption than a CCD image sensor, although this benefit is at the expense of the image quality. For microsatellites and nanosatellites where size and mass play a vital role, CMOS sensors are preferred, not least because they are cheaper (Fig. 12.9).

12.4.2 Earth/Horizon Sensors

The Earth provides a reference direction for determining the relative attitude of a spacecraft. Unlike the Sun, Earth

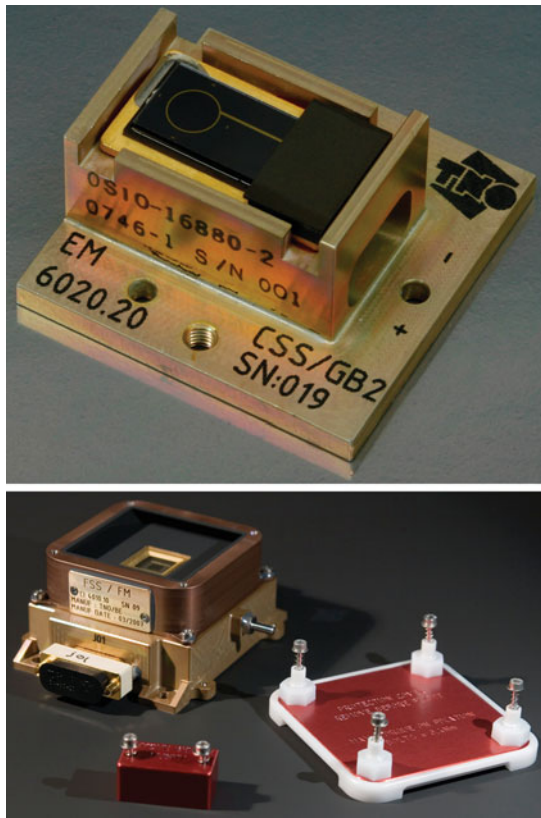


Fig. 12.8 Coarse analog cosine Sun sensor (*top*) and a fine, or two-axis analog cosine Sun sensor (*bottom*), both developed by Moog Bradford

cannot be approximated as a point source target, because in low Earth orbit around 40 % of a satellite's vision is filled up by the Earth. It is sufficiently easy to 'see' the Earth from the spacecraft through a wide range of spectral bands. However, in order for the sensor to differentiate Earth from cold space all along the orbit, the radiance emitted should have a uniform energy distribution over a range of the spectrum. The Earth's albedo lies mostly in the visible spectral range, and it varies widely depending on the reflecting surface (ice, snow, forest, water, soil, etc.) and on the time of the day/year, therefore causing ambiguity in the measurements. A spectral region that better matches the requirements is the infrared region. The spectral range of 14 to 16 μm (the CO_2 band) is used by most horizon sensors because the energy emitted has a uniform energy distribution irrespective of the day/night terminator and the Earth's reflecting surface as a result of most of the radiation being from the atmosphere above the Earth's surface [25, 26].

12.4.2.1 Principle of Operation of an Earth Sensor

An Earth sensor normally operates by scanning the sky to detect the Earth's horizon. It has an optical system detector, along with a signal-processing unit. If the Earth sensor is on

a spin-stabilized platform, then as the spacecraft spins the sensor can scan the sky to detect for infrared radiation emitted by the Earth. Otherwise, the sensor is provided with a steerable scanning mechanism in order to scan the sky. The basic components of an Earth horizon sensor are shown in Fig. 12.10.

The detector normally is a bolometer whose resistance varies depending on the incident radiation. Thereby, when the Earth is in the field of view of the sensor, the bolometer has a certain value of resistance and at other times the bolometer has a different resistance value. Other detectors that are useful include the photodiode (sensitivity mainly in the near-infrared region), pyroelectric devices, and a thermopile. The sensitivity of these detectors to the incident radiation primarily decides the accuracy of the Earth sensor. As the sensor works, basically, by detecting the horizon of the Earth it is also called the horizon sensor. The Earth sensor on-board the spacecraft traces out the base of a cone, Fig. 12.11 shows the geometry of the scan and the scan path on Earth.

The STD 15 EADS Sodern sensor is shown in Fig. 12.12, which is used to measure pitch and roll angles. It has a dual-track scanning pattern, hence rather than scanning a simple cone as shown in Fig. 12.11, it scans a shape akin to a cartoon bone, allowing it to measure pitch and roll angles at altitude between 15,000 and 140,000 km, with an operating nominal de-pointing pitch range of $\pm 12^\circ$ (roll = 0°) and a roll range of $\pm 2.9^\circ$ (pitch = 0°). With an accuracy budget 3σ , bias amounts to 0.035° and the typical noise to 0.015° [26].

12.4.3 Star Sensors

Star cameras can provide accurate, absolute attitude information by imaging stars and matching them to catalog positions. They provide the most accurate attitude information of all satellite sensors—an estimated accuracy of 20 arc seconds or less is typical. A star camera's performance depends on its ability to detect dim light sources, and the attitude accuracy improves with the number of stars that can be detected. CCD devices are usually used for imaging (as opposed to CMOS imaging devices) because they are more sensitive to the incoming photons. Star trackers have large apertures to allow as much light as possible to enter the lens, and make use of baffles to suppress stray light from the Sun or light reflected from the Earth or Moon. A star camera can be used in both the eclipse and daylight portions of the orbit as long as the bore-sight points away from these bright objects. The imaging devices used in star camera are susceptible to radiation effects. Radiation particles will typically damage pixels on the sensor, progressively reducing its star detection capabilities over time (Fig. 12.13).

Fig. 12.9 LISA Pathfinder digital Sun sensor flight model, including removable alignment cube (left). Image Galileo Avionica

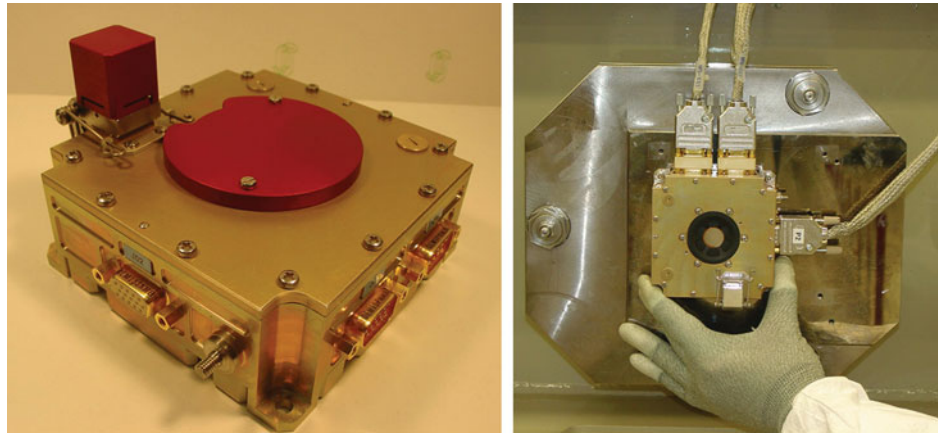


Fig. 12.10 Earth sensor components. Image Malcolm Macdonald

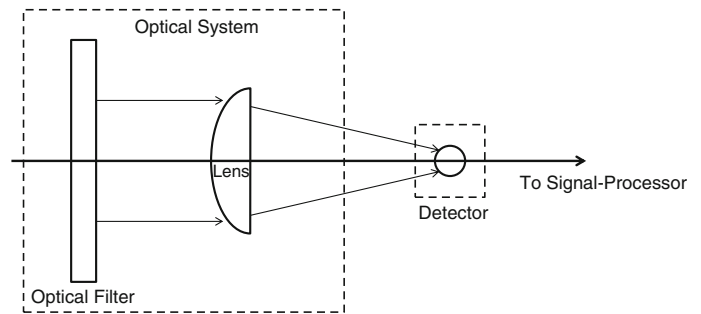


Fig. 12.11 Horizon crossing indicators. Left, scan geometry; Right, pulse generated. Image Malcolm Macdonald

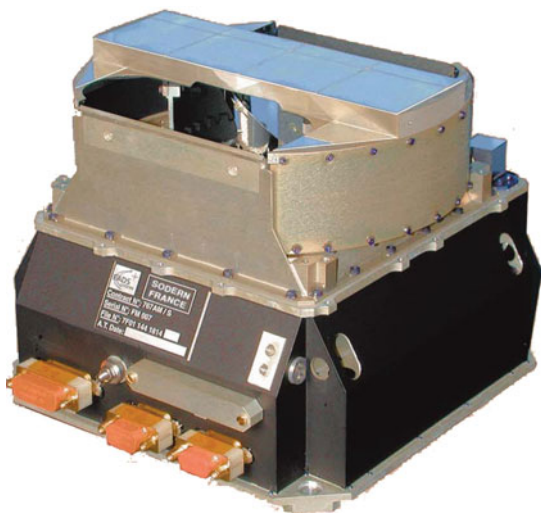
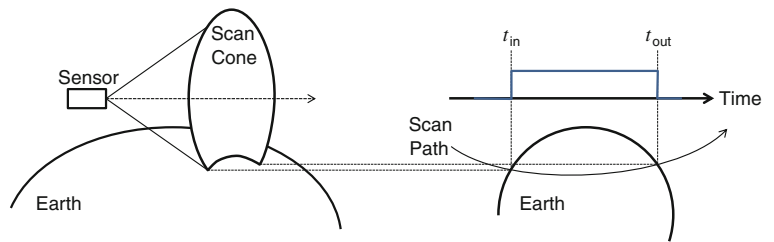


Fig. 12.12 The EADS Sodern STD 15 dual conical scanning Earth Sensor

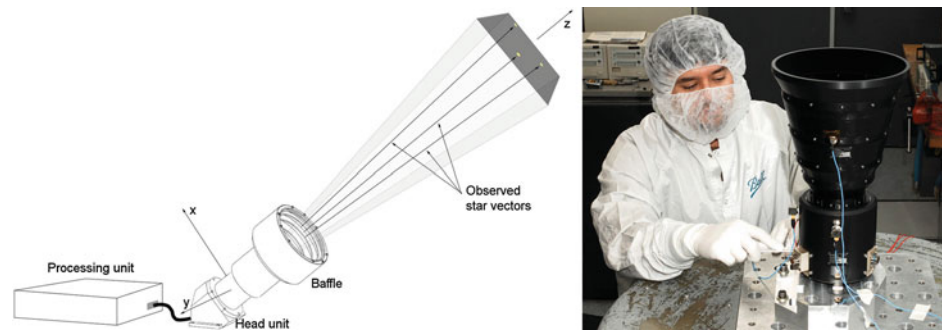
12.4.3.1 Attitude Determination from Matched Star Vectors

Reference star positions are obtained from a catalog, such as the Hipparcos and Bright Star Catalog (BSC). Detected stars are matched to catalog stars by finding pairs of stars with the same angular separation and the same apparent magnitude. A matching tolerance is used to allow for detection variations. Multiple pairs of catalog stars may be matched with a detected pair. These matches are then further pruned using a constellation matching algorithm [27].

12.4.3.2 Performance, Attitude Accuracy and Limitations

The number of stars that can be detected will depend on the camera's ability to discern brighter pixels from 'dark' pixels. Factors that influence this include the amount of light entering the optics and falling on the sensor, the noise characteristics of the sensor and readout electronics, the

Fig. 12.13 Star camera schematic (*left*) and the star tracker from the Kepler spacecraft (*right*). Image Ball Aerospace (*right*)



integration time of the sensor, and the suppression of stray light reflected in the optics. Star brightness is measured on a logarithmic scale, with negative numbers being very bright and larger positive numbers having less brightness. The human eye can discern stars as faint as magnitude 6.0 against a dark sky, but optical aid is needed for dimmer stars. Likewise, a star tracker will have a cut-off magnitude where dimmer stars are no longer detectable. For the star tracker to be able to operate for all possible orientations, it should be able to detect enough stars in its field of view for any given orientation. Light from bright objects like the Sun or Earth can influence the detection capabilities even if they do not appear in the star tracker's field of view. Stray light can still enter the optics at an angle and internal reflections can cause these rays to fall on the sensor, raising the average 'dark' level and effectively washing out the dimmer stars. Most star cameras have large baffles—cones with rings on the inside—that prevent stray light from entering the optics.

The star-magnitude detection threshold can be improved by allowing the sensor to integrate for longer. That way, more light is accumulated and a better signal-to-noise ratio is achieved. However, a longer integration time has its drawbacks. Firstly, a longer integration time implies a slower update rate. But more importantly, a longer integration time will only be helpful if the spacecraft is fixed with respect to the stars. A rotating star tracker will result in light from a single star 'smearing' over multiple pixels. Most star trackers will specify a maximum rotation rate at which it can operate (typically 5 degrees per second). For larger angular rates the smear will be spread out over too many pixels for detection to be possible.

The time that it takes a star camera to detect and match stars can be shortened if prior information about the attitude is known. For this reason, most star cameras incorporate a 'tracking' mode where previously detected and matched stars are tracked over small displacements. In the initial acquisition phase, when no attitude information is available, the entire catalog is searched for possible matches, but once matches have been established and an attitude estimate is available, this estimate can be used in subsequent iterations to limit the number of catalog stars that have to be searched. The result is a faster update rate. The attitude estimated by the star

tracker can be represented as an azimuth and elevation angle (X and Y), and a rotation around the camera's bore-sight (Z). Because all the measured vectors from the star tracker will be found inside a cone around the camera's bore-sight, the accuracy of the rotation angle around the bore-sight will be less than the azimuth and elevation accuracy. This effect will become more pronounced as the field of view is made smaller. To compensate for this, some star trackers have the ability to fit more than one camera head unit. The highest accuracy will be obtained if two camera head units are perpendicular. Adding more cameras will also increase redundancy in case of failure, and increase the availability of stars (in case one of the cameras is blinded by stray light).

12.4.4 Magnetometers

Magnetometers have become one of the most commonly used attitude determination sensors for satellites in LEO. This is primarily due to their simplicity, robustness, low cost, and small mass. They are used to measure the strength and direction of the local magnetic field. When this information is combined with a model of the Earth's magnetic field such as the International Geomagnetic Reference Field (IGRF) model, the attitude of the satellite can be determined. However, because the field is not well mapped and has many anomalies, magnetometers can only provide coarse measurements and they are usually combined with other sensors such as star cameras and Sun sensors. Magnetometers are extensively used in the de-tumbling phase of the satellite when magnetorquers are used for that purpose. The firing of the magnetorquers must be timed so as to allow the magnetic field to break down before any readings are taken. Special consideration needs to be taken as to the placement of the magnetometer. It must be placed away from any sources that might cause noise. For this reason, they are often placed at the end of extensible booms. The most common type of magnetometer for attitude determination purposes is the fluxgate magnetometer. These usually have a sensitivity of ± 10 nT with a range of ± 60 μ T. They tend to have an accuracy of between 0.5 and 5° and are only usable for altitudes below approximately 6,000 km [28].

12.4.5 Rate Gyros

Spinning gyros are one of the oldest and most popular sensors used on-board satellites and aerospace vehicles. Gyros can measure the angular rates of a vehicle without needing any knowledge of an external or absolute reference. Thus, if the attitude of a spacecraft is determined with an Earth or Sun sensor, then the angular rates of the satellite's principal axes can be obtained by differentiating the angular position outputs of the sensors. However, if the spacecraft enters an eclipse where this will no longer be possible, continued control will require the use of the gyro-rate sensors to measure the satellite's attitude. Another reason for using gyro-rate sensors is the need to control the angular rate of a spacecraft in addition to its angular position. Differentiating angular position outputs of a satellite from other sensors in order to get the angular rates can lead to noisy results, which will affect the stability and pointing of the ADCS system.

A gyro consists of a spinning wheel that reacts and measures imposed attitude rotations of a vehicle. The most common types of gyro are the rate-gyro (RG) and the rate-integrating gyro (RIG), which are based on the gyroscopic stiffness of revolving moments of inertia. The biggest disadvantage of a gyro is its reliance on moving parts, which have a limited lifetime. However, advances in mechanics, microelectronics and space components have contributed to the development of sensors based on new physical concepts involving no moving parts, such as laser gyros, quartz rate sensors, MEMS sensors, fiber optic gyros (FOG), and hemispherical resonator gyros (HRG).

In order to be able to understand and compare rate gyros which come with inherit noise problems, various design and performance parameters are defined

- *Range*—The larger the range of measurement, the larger the noise level of the sensor. The smaller the range, the better the accuracy; ranges of 1–100°/s are feasible.
- *Bias (drift)*—The most important characteristic of a gyro and intrinsic to the technology, ranging from 0.01 to 1°/hr.
- *Output noise*—Specified per frequency band.
- *Scale factor*—Important for rate integration and has a strong influence on the achievable attitude accuracy.

Gyros are commonly used in clusters, one per axis plus a fourth unit in a skewed configuration for redundancy, but it is also possible to have all four off-axis in order to maximize redundancy. This particular configuration is also called an inertial reference unit (IRU). The combination of gyros and accelerometers can give addition position/velocity measurements, and this is called an inertial measurement unit (IMU).

12.4.6 Global Navigation Satellite System/ Global Positioning System

Global navigation satellite system (GNSS)/Global Positioning System (GPS) signals can be used both for orbit determination (position) and for attitude determination using multiple antenna layouts. A set of antennas is used, connected to a GPS receiver on the top panel of a spacecraft in LEO, facing the GNSS constellations in MEO. Using the phase difference between the antennas allows a reconstruction of the attitude of the spacecraft. Despite various technical issues such as multipath and noise, accuracies of 0.1° to 1° have been achieved [29, 30].

12.5 Attitude Control Actuators

Actuators can be divided into inertial and non-inertial types. Inertial actuators are devices that generate torques, by modifying their angular momentum. They include

- *Momentum wheels (MW)*—They provide constant angular momentum for gyroscopic stabilization. Orientation of the spin axis is fixed with respect to inertial space. Attitude control is achieved by varying the spin speed of the wheel about some nominal value.
- *Reaction wheels (RW)*—They provide torque to a vehicle by increasing or decreasing the speed of the wheel, with the wheel nominally at rest.
- *Control moment gyroscopes (CMG)*—A momentum wheel gimbaled in one or two axes. Control torques are generated by changing the direction of the spinning wheel's axis to vary the direction of the momentum vector.

Non-inertial actuators are

- *Magnetic torquers (MT)*—Magnetic coils or electromagnets that generate magnetic dipole moments. A magnetic torquer produces a torque proportional (and perpendicular) to the Earth's magnetic field. It is often used as a secondary actuator on a spacecraft in order to de-saturate momentum exchange systems, although it has become a common primary system on CubeSats and other resource-limited spacecraft.
- *Thrusters*—Produce a thrust (force) or torque around the center of mass by expelling propellant.

Another means of applying attitude control actuation is to manipulate the attitude disturbance torques in a favorable way. The most common of these techniques is to use solar radiation pressure for attitude control on GEO platforms, where a reflective trim tab is often used to aid attitude control and reduce propellant consumption. It is noteworthy that the same method was used on the Mariner-10 spacecraft to Mercury and Venus, and on the MESSENGER

Fig. 12.14 Reaction/momentum wheels. Rockwell Collins Teldix space wheel shown without outer casing (*top left*), the reaction wheel from the Kepler spacecraft (*bottom left*) and one of X-ray Multi-Mirror (XMM)-Newton's four reaction wheels. *Image* Ball Aerospace (*bottom left*) and Matra Marconi Space, UK (*right*)



probe to Mercury, which used solar radiation pressure on their solar panels to perform fine trajectory corrections. The Japanese Hayabusa spacecraft also used solar radiation pressure for attitude control in a recovery mode after the failure of its on-board reaction wheels. The manipulation of attitude disturbance torques to aid attitude control will not be discussed further here.

12.5.1 Momentum/Reaction Wheels

Momentum/reaction wheel (MW/RW) systems operate on the principle of conservation of angular momentum. Using rotating masses in a spacecraft's body allows the transfer of angular momentum between different parts of the vehicle without changing its overall angular momentum. Inside a spacecraft, a symmetrical rotating body produces angular torque when accelerated about its axis of rotation. The rotating body may have an initial constant angular momentum (spinning flywheel). As this momentum is internal to the spacecraft, its increase does not change the total momentum of the system but rather transfers the momentum to the spacecraft.

The ratio between the satellite and flywheel inertia is selected such that it fulfills the attitude control (agility) requirements of a specific mission, taking into consideration, mass, volume and power constraints for the actuators to be used on a satellite. Sections 12.1.2 and 12.1.5 detail the mathematical models used for spacecraft using attitude control actuators.

A view of typical wheels is shown in Fig. 12.14. They consist of a precision engineered flywheel with most of the mass concentrated at the tip/rim of the disk to achieve maximum wheel inertia for a given mass. A brushless DC motor is usually used to rotate the wheel. The complete wheel assembly with integrated electronics is housed in a cage. This cage has a two-fold task; it helps to protect the

spacecraft in case something goes wrong with the spinning wheel, and it also sometimes acts as a pressure vessel to keep the lubricant from outgassing. The bearing assembly, which is required for mechanical support and operations, is what limits the lifetime of reaction wheels to about 5–15 years, depending on the duty cycle of the wheel, the type of lubrication used, and the motor technology. The design of miniature, low-cost and low-jitter MW/RW with longer lifetimes has continued in parallel with new mechanical and mechatronic developments.

Reaction wheels are used when accurate time-optimal rapid maneuvers are required. They allow continuous and smooth control of torque, can accelerate in both directions, and normally have a zero speed. However, due to friction they display a non-linear response at very low spin rates that might cause an irregular motion of the spacecraft. This is usually solved by running the wheel with a small bias.

Momentum wheels are essentially the same as reaction wheels but have a large nominal spin rate. This provides a constant angular momentum that causes gyroscopic stiffness around two axes that helps to maintain the attitude of the spacecraft.

As previously stated, a minimum of three non-coplanar wheels are required for full three axis control. In order to avoid single-point failure, a fourth wheel is usually added in a skewed configuration, but it is also possible to have all four off-axis in order to maximize redundancy. Due to the addition of a fourth wheel, which is usually added at an equal angle with the other three wheels, additional torque and momentum authority may be required.

The wheels need to be properly sized so as not to become saturated by the expected worst-case disturbance torques. When momentum builds up, external torque actuators such as magnetorquers or thrusters are required to dump some of the momentum. The torque capability of the reaction wheels is determined by the desired slew rates.

12.5.2 CMG: Control Moment Gyros

Control moment gyros (CMG) are gimballed wheels that can generate large amounts of torque/angular momentum. They are considered to be ‘torque amplifiers’ because they can use the stored angular momentum in a flywheel and ‘convert’ it to large torques by gimbaling the flywheel appropriately.

A CMG consists of two parts

- The momentum wheel, which produces a large and constant angular momentum (magnitude).
- The gimbal motor (or set of gimbal motors) on which the momentum wheel is mounted, so that the angular momentum vector of the wheel can be changed to the desired direction.

As shown in Fig. 12.15, torquing the gimbal results in a precessional torque that is normal to the gimbal axis and spin axis of the momentum wheel

$$\mathbf{N}_{\text{CMG}} = \mathbf{h} \times \dot{\boldsymbol{\delta}} \quad (12.143)$$

where \mathbf{h} is the angular momentum vector and $\dot{\boldsymbol{\delta}}$ is the gimbal rate.

Depending on the mechanical characteristics, CMGs can be characterized as

- *Single-gimbal CMGs (SGCMG)*—The momentum wheel is gimballed in one axis and constrained to rotate on a circle in a plane that is normal to the gimbal axis.
- *Double-gimbal CMGs (DGCMG)*—The momentum wheel is constrained inside two gimbals and the angular momentum vector is oriented within a sphere.
- *Variable-speed CMGs: SGCMG*—A variable speed momentum wheel provides an extra degree of control than is available to SGCMGs.

The advantages and disadvantages of each is summarized in Table 12.2.

Most CMGs are used on large spacecraft, principally due to their high angular momentum storage capability, which provides increased stabilization under large external disturbance torques. CMGs can also produce substantial torques and are very heavy (typically 55–150 kg for 300–1,000 Nms momenta and 100–1,000 Nm output torques [8]). Recently, Astrium and Honeywell have begun to work on smaller CMGs (see Table 12.3), for new families of spacecraft in the 500–2,000 kg range that require high precision pointing and fast slew capabilities [5, 31, 32].

The Skylab and Mir space stations both used CMGs for attitude control and stability, as does the International Space Station (ISS). Skylab used a cluster of DGCMGs, while Mir had a cluster of six SGCMGs with only four being used at a time. The ISS uses four DGCMGs in a boxed configuration. One of the ISS CMGs suffered a bearing failure in 2002, leaving the space station with two primary gyroscopes and one spare. That spare shut down in 2004 when a circuit

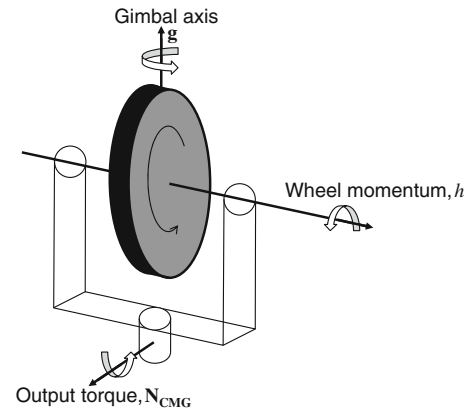


Fig. 12.15 Single-gimbal CMG diagram

breaker failed. Though subsequently repaired, it failed again in March 2005. When power was rerouted during a spacewalk, this restored it to operation [33] (Fig. 12.16).

Honeywell is the biggest manufacturer of CMGs and they are used in a large number of spacecraft [5, 34]. The company has developed the M-50 CMG for agile satellites in the 1,000–2,000 kg class for Earth observation [34]. The M-50 has flown on two spacecraft, the Ball Aerospace Worldview-1 and -2, together with Quickbird from DigitalGlobe’s constellation of commercial remote sensing satellites. The more recent spacecraft, Worldview-2 was launched in 2009. It is a very agile platform providing re-targeting capability with an acceleration of $1.5/s^2$, with a slew rate of $3.5/s$ and a time to slew 200 km is 10 s for a 2,800 kg spacecraft with 0.5 m panchromatic, 2 m multi-spectral imaging capability.

Astrium is also building a CMG based on a Teldix RW. The compact CMG has been designed for the French Pleiades spacecraft, currently in orbit and a new platform in the 1 ton class designed for agile, high resolution imaging in a constellation of imaging (provided by France) and radar (provided by Italy) spacecraft [35].

On the other end of the performance spectrum, a twin micro-CMG payload was designed and built by the University of Surrey and SSTL for the Turkish BILSAT-1 imaging microsatellite flown in 2002, demonstrating that agility with the implementation of low cost micro-CMGs based on COTS is feasible [8, 31, 36].

12.5.3 Magnetorquers/Magnetic Control

Magnetic control has been used in many space missions. The simplicity, inexpensive hardware and reasonably good attitude control (0.5° to 5° in all axes) makes magnetic control very attractive, especially for small satellites primarily for attitude control and momentum dumping of reaction/momentum wheels on small satellites.

Table 12.2 Advantages and disadvantages of different CMG types

CMG type	Advantage	Disadvantage
SGCMG	Great torque amplification	Singularities
DGCMG	Torque amplification, extra degree of freedom	Cost, complexity, size
VSCMG	Extra degree of control	Enhanced control

Table 12.3 CMGs with flight heritage

Year	Manufacturer	Type of CMG	Torque	Momentum	Mass (Kg)
1973	Bendix (Skylab)	3 DGCMGs	N/A	2,700 Nms	200
1976	Honeywell	4-6 SGCMGs	>150 Nm	>100 Nm	>30
1986	MIR	4-6 SGCMGs	N/A	N/A	N/A
2001	L-3 (ISS)	4 DGCMG	258 Nm	4,760 Nms	272
1999	Honeywell (M50)	SGCMG	85 Nm	25–75 Nms	28
1999	Pleiades (Astrium)	SGCMG	45 Nm	15 Nms	15.7
2002	SSC/SSTL	SGCMG	50 mNm	0.28 Nms	2
2007/2009	WorldView-1 & 2	SGCMG	85 Nm	25–75	28

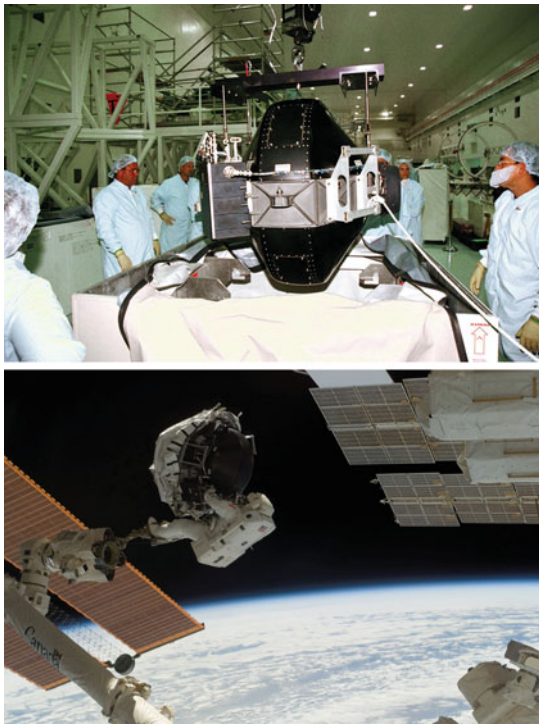


Fig. 12.16 Boeing technicians remove the cover from a Control Moment Gyroscope (CMG) in the Space Station Processing Facility at Kennedy Space Centre (*top*) and astronaut Dave Williams, STS-118 mission specialist, anchored to the foot restraint on the Canadarm2 removing a faulty control moment gyroscope (CMG-3) and installing a new CMG into the station's Z1 truss (*bottom*). *Image* NASA; bottom S118-E-06993 (13 August 2007)

Interaction between a magnetic moment, \mathbf{M} , generated by a spacecraft with the Earth's magnetic field, \mathbf{B} , produces a control torque \mathbf{N}_M acting on the spacecraft

$$\mathbf{N}_M = \mathbf{M} \times \mathbf{B}. \quad (12.144)$$

The direction of \mathbf{M} can be controlled on average by a proper sequence of magnetorquers firings, but the \mathbf{B} field vector is dependent on the orbital location. As a result, the torque \mathbf{N}_M , which always is orthogonal to \mathbf{B} and \mathbf{M} , is not necessarily favorable for control of the attitude of a specific spacecraft axis in certain regions of the orbit. Another drawback of magnetorquers is that it is possible that a desirable control torque for a certain attitude axis (pitch, roll, yaw) might generate undesirable disturbance torques for the other axes. The Earth's magnetic field is predominately a magnetic dipole. The magnetic field can be expressed mathematically by a spherical harmonic model, the so-called IGRF (International Geomagnetic Reference Field) model. For purposes of simulation, a first-order dipole model is often used to represent the geomagnetic field vector. This dipole vector is expressed as

$$\mathbf{B} = \nabla \left[\frac{\mathbf{R}^T \mathbf{M}_e}{R_s^3} \right] = [\mathbf{1} - 3\mathbf{R}\mathbf{R}^T] \frac{\mathbf{M}_e}{R_s^3} \quad (12.145)$$

where ∇ is the vector gradient operator, R_s is the length of the geocentric position vector, \mathbf{R} is the unit geocentric position vector, \mathbf{M}_e is the geomagnetic strength of the dipole vector, and $\mathbf{1}$ is the identity matrix.

Magnetorquers can produce a torque based on the magnetic moment that it can produce, which depends on the number of coil windings n , the cross-sectional area A of the coil, and on the amount of current I that passes through the coil in the unit vector along the coil's axis \mathbf{u}

$$\mathbf{M} = nI\mathbf{A}\mathbf{u}. \quad (12.146)$$

Combining Eqs. 12.144 and 12.145 gives

$$\mathbf{N}_M = nIA(\mathbf{u} \times \mathbf{B}). \quad (12.147)$$

Usually, three magnetorquers are used on a spacecraft, one per axis for coarse attitude control and mainly for angular momentum unloading. Their utility decreases with increasingly altitude due to the decreasing strength of the magnetic field. The field's strength and direction also varies. A specific feature of magnetorquers is that they cannot produce a torque component about the local field direction. For example, in a polar orbit any required direction can always be achieved at some point in the orbit since the field direction changes round the orbit, whereas in the equatorial plane this would be problematic because the field lines are always in a north-south direction. Magnetorquers do not require any propellant, require very limited power levels, and have an unlimited lifetime, as well as no moving parts. Therefore, they are very popular, simple to manufacture and inexpensive actuators.

12.5.4 Thrusters

External disturbances acting on satellite can be countered by using small thrusters, thus controlling the total momentum of the spacecraft. They are mounted in clusters on the surfaces of a satellite in various configurations in order to provide the required direction of torque about each axis. The disadvantages of using a thruster, especially when compared to magnetorquers is the consumption of propellant, increased mass, complexity and cost. However, using thrusters is independent of altitude and of the Earth's magnetic field, and they can be used for fine/precise attitude control, station-keeping of GEO satellites and as an orbit control system in many cases.

Europe's Automated Transfer Vehicle (ATV) built by EADS Astrium for ESA is an automated robotic tug that transports cargo to the ISS using an autonomous ranging and rendezvous system which requires very fine and precise attitude control. The ATV, which has a mass of 20,750 kg at launch, and uses twenty-eight 220 N bipropellant thrusters for attitude control. Figure 12.17 shows the ATV with four clusters of two thrusters and four clusters of five thrusters [37].

12.5.5 ADCS Heritage Design Case Studies

12.5.5.1 Earth Observation Small Satellites

UoSAT-12 is a low-cost minisatellite built by Surrey Satellite Technology Ltd. (SSTL), and amongst other

objectives it is a technology demonstrator for high performance attitude control and orbit maintenance on a future constellation of Earth observation satellites. The satellite uses a 3-axis reaction wheel configuration and a cold gas propulsion system to enable precise and fast control of its attitude, for example during orbit maneuvers. Magnetorquer coils assist the wheels mainly for momentum dumping. This section describes the various attitude control modes required to support: (1) the initial attitude acquisition phase, (2) a high-resolution imager payload during pointing and tracking of targets, and (3) the propulsion system during orbit maneuvers.

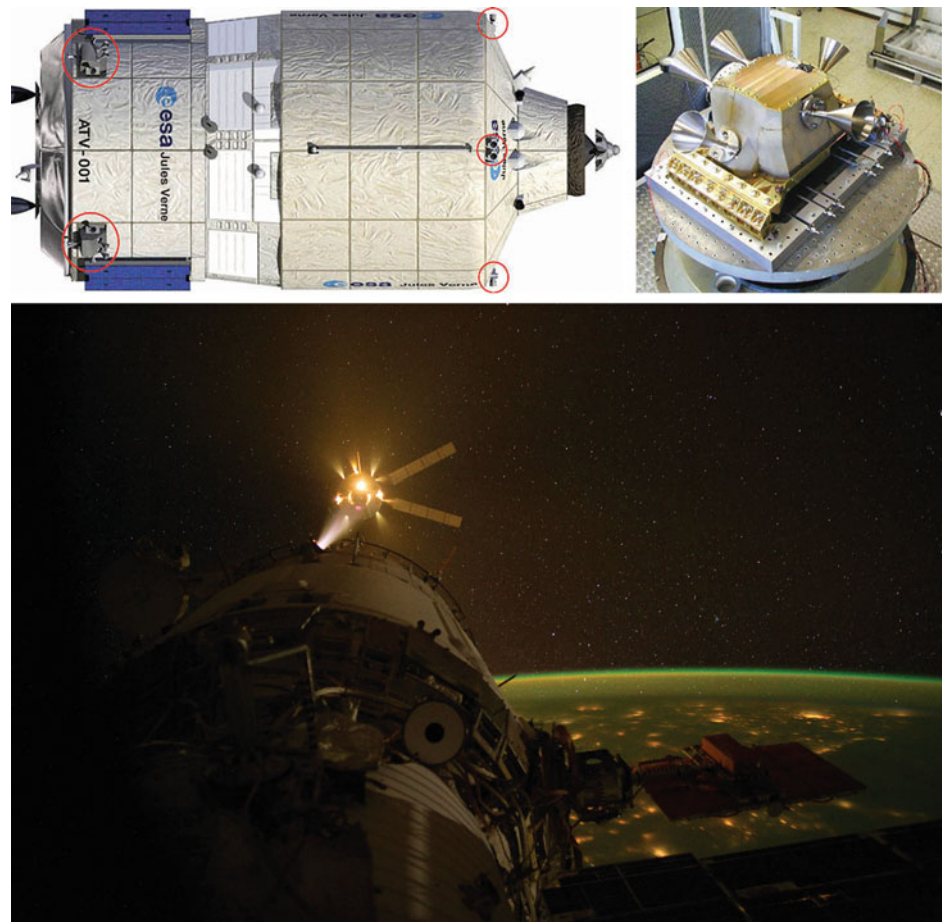
UoSAT-12 supports a wide range of sensors for attitude determination and a multi-channel GPS receiver for on-board orbit determination and accurate time synchronization. The GPS receiver also has an experimental attitude determination capability, through baseline sensing of an array of five patch antennas. A redundant set of three 3-axis fluxgate magnetometers are used to measure the geomagnetic field vector in the satellite's body coordinates. These measurements are used to determine the magnetic coil torque vector and in combination with a magnetic field model to estimate the full attitude and angular rates of the satellite. Four 2-axis (azimuth and elevation) Sun sensors measure the Sun vector angle to high accuracy. During the nominal nadir pointing attitude mode, small pitch and roll angles can be measured with a 2-axis infrared horizon sensor. The highest attitude measurement accuracy is obtained from a dual set of opposite looking star sensors. The sensors supply star measurement vectors and matched star catalog vectors at a rate of once per second to an attitude and rate estimation filter. A solid-state angular rate sensor is mounted in one axis of the satellite to flight qualify the sensor for future missions. Table 12.4 lists all the sensors used on UoSAT-12 for attitude determination and their most important characteristics.

Twelve magnetorquer coils are positioned within the satellite to give some level of redundancy and to deliver full 3-axis magnetic dipole moment control. These coils are controlled using dual polarity current pulse width control in order to deliver the required averaged level of magnetic moment per sample period. The magnetorquers are used for

- De-tumbling of body angular rates after launch
- Momentum dumping of the reaction wheels
- Momentum maintenance on the momentum wheel during Y-spin stabilization
- Nutation damping during spin stabilization
- Libration damping and yaw spin/phase control after deployment of a backup gravity gradient boom.

Three momentum/reaction wheel subassemblies are mounted in a 3-axis configuration to enable full control of the attitude or angular momentum of the satellite. One wheel is a space qualified wheel from Ithaco and is mounted

Fig. 12.17 Automated transfer vehicle (ATV), with 220 N thruster clusters shown forward and aft (*top left*) and a five thruster cluster (*top right*). ATV Edoardo Amaldi approaching the ISS for docking on March 28, 2012 (*bottom*); taken by NASA astronaut Don Pettit on-board the ISS, with the ATV thrusters firing under automated control as the vessel nears the Russian module where it docked. *Image* ESA (*top left and right*) and NASA (*bottom*)



in the structural Y-axis direction in order to have higher reliability when the pitch momentum bias control mode is used. The-SSTL manufactured wheels, destined to be space qualified on UoSAT-12, are mounted in the structural X and Z-axis directions. The wheels are used for the following control functions

- Full 3-axis pointing and slow slew maneuvers during imaging
- Nadir, Sun or inertial pointing of the payloads by using angular momentum stiffening
- Near minimum-time Euler-axis rotations for quick attitude maneuvers
- Fast spin-up or spin-down of the satellite body e.g. barbecue mode of the solar arrays
- Cancellation of the disturbance torque caused by the propulsion system during orbit control
- Thruster and moment of inertia calibration.

UoSAT-12 is fitted with a single nitrous oxide (N_2O) resistojet thruster (see [Chap. 11](#)) for orbit maintenance and ten cold gas nitrogen thrusters for orbit or attitude control. The resistojet is aligned to the center of mass of the satellite. Some cold gas thrusters can be used in pairs to limit the

attitude disturbance torque during orbit control maneuvers. The propulsion system can be used for the following functions

- Full 3-axis rough pointing and fast slew control
- Drag compensation of the satellite's orbit
- Orbit shaping to demonstrate constellation control
- Wheel momentum dumping/maintenance.

Table 12.5 lists the various actuators used on UoSAT-12 for attitude and orbit control and their respective characteristics.

UoSAT-12 was launched on April 21, 1999, from the Baikonur Cosmodrome in Kazakhstan into a 650 km circular orbit at 65° inclination. The initial telemetry of the magnetometer indicated a tumbling rate of about $2^\circ/s$. The next day the ADCS software was loaded on the 186-OBC (on-board computer) and the Rate Kalman filter confirmed the initial tumbling rate to be mainly around the Y-axis. This result confirmed the slightly higher Y-axis moment of inertia and the cross-products of inertia that had been predicted pre-launch. Initially only the Y-axis magnetorquer was used to dump the X and Z-axis angular rates, then the X-axis magnetorquer was enabled to control the Y-axis rate

Table 12.4 Attitude and orbit determination sensors on UoSAT-12

	Magnetometer	Sun sensor	Horizon sensor	Star sensor	Rate gyro	GPS
Manufacturer	SSTL (2) & Ultra (1)	SSTL	Servo—MiDES SSTL I/F	SSTL	BEI SSTL IF	SSTL
Quantity	3 units	4, 2-axis	1, 2-axis	2 units	1 unit	1 unit
Type	Fluxgate	Slit and photo cell	IR pyro array & chopper	CCD matrix	Gyrochip	Mitel chip set, 24 channels, 4 antennae
Range	$\pm 60 \mu$ tesla	$\pm 50^\circ$	$\pm 5.5^\circ$	$14.4^\circ \times 19.2^\circ$	$\pm 5^\circ/s$	
Accuracy	30 n tesla (3σ)	0.2° (3σ)	0.06° (3σ)	0.02° (3σ)	$0.02^\circ/s$	50 m (1σ)
Power	<0.8 W	<0.1 W	2.8 W	4 W	1.4 W	5–7 W

Table 12.5 Attitude and orbit control actuators on UoSAT-12

	Magnetotorquers	Reaction/momentum wheels	Propulsion system
Manufacturer	SSTL	SSTL (2); Ithaco (1)	SSTL & Polyflex
Quantity	8 \times PCB 4 \times Wire coils	3 units (Z, Y, Z)	10 \times N ₂ CG thrusters 1 \times N ₂ O resistojets
Type	Air Core	Brushless DC motor, dry lubricated bearings	4 \times bar cold gas N ₂ O plus 100 W heater
Operation range	X/Y = $\pm 14.2 \text{ Am}^2$ Z = $\pm 13.3 \text{ Am}^2$	$\pm 4 \text{ Nms}$ @ $\pm 5,000 \text{ rpm}$ $\pm 0.02 \text{ Nm max}$	Thrust: 0.1 N (CG) & 0.125 N (R-jet) ΔV : 14 m/s (CG) & 9.7 m/s (R-jet)
Power	20 W max, and 80 % duty cycle	2.8–14.6 W (zero to max. acceleration)	3 W (CG) 100 W (R-jet)
Operation	PMW controlled	Speed controlled	PMW controlled
Accuracy	20 ms minimum pulse	$\pm 1 \text{ rpm}$	>10 ms pulse (CG) >600 s pulse (R-jet)

towards the Y-Thomson reference rate of $-1^\circ/s$. After two orbits, the body angular momentum of UoSAT-12 was almost completely dumped. The magnetorquer controller was left running for two more orbits until the satellite was in the required Y-Thomson attitude. The cross-products of inertia prevented the satellite from reaching the target rate exactly, and the true Y-rate estimated was approximately $-0.8^\circ/s$ with small residual X and Z-rates of less than $0.2^\circ/s$.

Following transition to nadir-pointing, the ADCS system was put in zero momentum mode. The pitch and roll pointing errors experienced were very small, the $1-\sigma$ deviation is 0.13° and the maximum peaks during this period are less than 0.5° . The yaw error was worse ($1-\sigma$ deviation is 0.62°) due to a lack of accurate yaw information close to the polar region and the use of the magnetometer as the only source of yaw information to the EKF estimator. Yaw control peaks of 3° were experienced for short periods at the maximum latitude extremes of each orbit. The reaction wheels were running mostly below 20 rpm and the magnetorquer peaks were 2.5 % of the saturation limit. The magnetorquers are being used exclusively to dump any wheel momentum build-up.

12.5.5.2 GEO Satellite Case Study

The Eurostar family has been a very successful platform for GEO satellites built by EADS Astrium. This section presents the architecture of the Eurostar 3000 AOCS system flown on the CNES Stentor (Satellite de Télécommunications pour Expérimenter les Nouvelles Technologies en Orbite), Amamzonas-2, Arabsat 5A/C, Interlsat-10, Hotbird 10, YahSat 1A/1B missions and many others (Fig. 12.18).

The Eurostar 3000 AOCS was designed by Matra Marconi Space (now Astrium) based on the successful Eurostar 2000+ AOCS concept and hardware with a focus on performance enhancement and reducing operational workload. The AOCS system was designed to be scalable for medium to very large GEO platforms. The AOCS system was validated in the CNES STENTOR mission flown in 2002. Heritage for the Eurostar 3000 AOCS system was also drawn from the Eurostar 1000/2000 AOCS, including the use of solar radiation pressure for smooth roll/yaw attitude control through the modulation of solar radiation pressure using fixed patented flaps on the solar arrays. With the use of electric propulsion (EP) the AOCS system achieved improved pointing stability. Yaw sensing becomes



Fig. 12.18 Astrium Eurostar 3000 Meosat-3b

important in EP station-keeping maneuvers, and this is achieved using solar array Sun sensors in a gyroless control mode.

The Eurostar 3000 AOCS subsystem uses a centralized computer concept. The AOCS system uses a four-wheel skewed configuration composed of two momentum wheels and two reaction wheels. The chemical propulsion subsystem uses fourteen, 10 N thrusters in two branches and a liquid apogee engine. The plasma propulsion subsystem is composed of two small platforms with plasma thrusters used for north/south station-keeping maneuvers and orbit eccentricity correction. The thruster direction is controlled by commanding stepper motors of the 2-axis thruster orientation mechanisms. The solar arrays equipped with flaps produce long-term inertial torques through the offset of each wing with respect to the Sun direction.

The AOCS subsystem is based on full redundancy. A hierarchical failure, detection, isolation and recovery (FDIR) is used to maintain the telecommunications mission in case of anomaly while limiting ground intervention.

The transfer and acquisition phase which is inherited from the Eurostar 2000+ sequence is based on 3-axis control. For the Eurostar 3000 an improved design is utilized allowing for simpler operations without any loss on safety. In the transfer phase, in which solar arrays are partially deployed, all operations from launcher separation to the end of the apogee sequence are included. In this phase only Sun and Earth acquisition is used to reach 3-axis pointing. The pointing information is then used for gyro stabilization in order to perform the LAE firing, but does not require any knowledge of the gyro scale factors.

When on-station, the attitude is based on a wheel system with two degrees of freedom, which deviates from the momentum bias system used on Eurostar 2000+. Here the solar radiation pressure mode is used exclusively for long-term wheel unloading. For large GEO satellites, it is not easy to separate the short-term/long-term movements due to

the nutation frequency. Thus, a robust momentum/control design is implemented for global large band control extending the bandwidth of the control.

Another innovation of the Eurostar 3000 AOCS system was the ability to provide gyroless yaw estimation during the EP station-keeping phase based on Earth and Sun sensor measurements.

12.6 Orbital Guidance, Navigation and Control Systems

This section is devoted to the guidance, navigation, and control (GNC) of the motion of the spacecraft center of mass. The focus is on GNC systems for on-board autonomous operations. The analogous problem for the spacecraft orientation was described in the previous sections. The typical functional architecture of the AOCS presented in Fig. 12.1 assumes separate navigation chains for determination of attitude and orbital motion. This architecture shows the typical data flow within the GNC system and its interfaces with the other functions of the AOCS.

The decoupling of GNC and ADCS is quite convenient for analysis and design. Some examples of missions with separated GNC and ADCS systems are NASA's Deep Space-1 [1] and Deep Impact [2] and JAXA's Hayabusa [3]. The interactions between ADCS and GNC will be presented in the application examples. It is worth noting that in some references (often dealing with highly autonomous systems) the orbit and attitude control system is referred to as the GNC system.

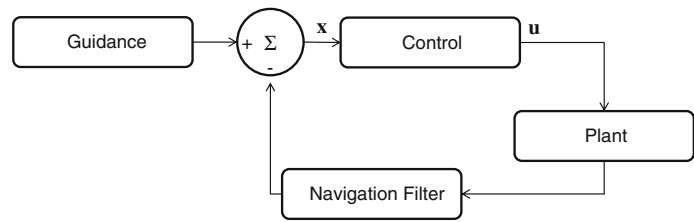
The on-board GNC system is responsible of the following tasks [4, 5]

- *Navigation* determines the present state with the required accuracy.
- *Guidance* creates the reference path to achieve the desired goal in nominal conditions.
- *Control* produces forces required to follow the reference path.

Expanding the GNC block of Fig. 12.1, a schematic diagram of a closed-loop control system is depicted in Fig. 12.19. The plant includes the actuators, the real world dynamics, and the sensors.

The GNC system must provide information to the top-level failure detection, isolation and recovery (FDIR) function and to the mission and vehicle management (MVM) function. The FDIR detects system and equipment failures and recovers from them. The modern paradigm requires FDIR functions at all levels. Thus, failure detection algorithms are included in the GNC functions. The MVM manages all the subsystems of the spacecraft, e.g. thermal, power. It defines the GNC modes and the sequencing of maneuvers. The measurement management function takes

Fig. 12.19 Schematic closed-loop control system



the raw observations from the sensors and produces validated measurements in the proper format for the navigation filter. Of particular relevance is the processing of raw images from an optical camera or an imaging LIDAR. This task is usually computationally expensive and is performed in a dedicated processor. The actuator management function takes the control output and issues the commands to the actuators. As these functions are not specific to the GNC they will not be detailed in this section.

The on-board GNC system must not be confused with the Flight Dynamics System that is part of the ground control system and will be discussed in Sect. 20.1.4, even though the objectives of both systems are quite similar. The GNC has tight constraints on time response (reactivity or responsiveness) and computational load. Thus, the performance of the on-board navigation and guidance functions can be worse than the ground-based analogs.

The GNC system must allow monitoring by the ground operators. In addition, the on-board system must permit updating of the GNC parameters using the ground information, including reset of the navigation function or change of the reference path.

The requirements of the GNC system depend on the level of autonomy of the spacecraft. If the position and motion are fully controlled from the ground, then the complete GNC system is not required. On-board autonomy aims to provide the space segment with the capability to continue mission operations and to survive critical situations without relying on ground segment intervention [6]. The on-board autonomy depends on the specific mission requirements and constraints. The autonomy level can vary between a very low level, involving a high level of control from ground, to a high level, whereby most of the functions are performed on-board.

The autonomy of the space segment has an impact on total life cycle cost. Increased autonomy can increase the development costs, but decreases the operating costs. Therefore, the adoption of specific autonomy goals for a given mission is decided by careful balancing of costs, risks, and schedules for both the development and the operation & maintenance phases. The need for autonomy is very different from one mission to another, and there are three factors that have a strong influence on the degree of autonomy that is required

- Communication delays, when the characteristic time of the mission is much shorter than ground control response time (including communication signal round-trip time).
- Environment uncertainty, where the safety of the mission demands high reactivity to unknown disturbances, for instance in missions to near-Earth objects (NEO).
- Costs and operation teams downsizing, which is very important during long routine phases.

In summary, autonomous GNC systems are complex and critical for space missions (in terms of cost, risk, and schedule). Therefore, nowadays autonomous GNC systems are implemented when they are the only feasible option, or in technology demonstration missions. Some examples where mission feasibility depends on autonomous GNC are robotic rendezvous in orbit around Mars (e.g. Mars Sample Return), pinpoint entry, descent and landing (EDL) for a solar system body, or a hypervelocity impact or flyby of a small body (e.g. Deep Impact). There have been advancements to reduce the operational costs of aerobraking at Mars by means of increased level of GNC autonomy (e.g. Mars Reconnaissance Orbiter [7]) which takes several months and has many uncertainties and risks.

Some examples of technology demonstration missions implementing autonomous GNC are Deep Space-1, in which the autonomous flight of long interplanetary low thrust arcs reduced operational costs, or Prisma [38] for rendezvous and formation flying.

12.6.1 Drivers for GNC Design

As follows from previous paragraphs, the capabilities and elements of the GNC system have to be designed on a case-by-case basis. Each mission has different objectives and constraints that will drive the GNC requirements. The requirements on the GNC system are derived from mission, system, or operational constraints. For instance, the GNC system required for Mars landing changes dramatically depending upon whether the landing dispersion is several hundred kilometers or a few kilometers.

Typical mission constraints are

- Phase goal, such as the final position dispersion in a rendezvous (between chaser and target), in a landing

(between desired and actual landing sites), or the final orbit when aerobraking.

- Safety issues, such as passive safe trajectories during the approach phase of a rendezvous or in formation flying, the safety corridor during the terminal approach of a rendezvous leading to mating (docking, berthing or capture).
- Initial conditions and environment model parameters, including uncertainties.
- Communications during critical phases, for instance sending vital information during EDL or monitoring during terminal rendezvous in Earth orbit.
- Overall cost, including algorithms and software development, equipment procurement, assembly, integration and validation (AIV) operations.
- Technology readiness level (TRL), selection of components with minimum TRL at given date.
Some examples of system constraints are
- Mass and size limits, considering redundant units.
- On-board resources such as propellant, computational load, memory, power, and thermal budget.
- Cost, including development and qualification (if necessary), AIV models, spare units.
- Location of equipment to avoid measurement degradation (interference, multipath, dazzling, shadowing, occultation, pollution, etc.).
- Constraints from other subsystems such as pointing accuracy (a.k.a. APE) or pointing stability (a.k.a. RPE).
Typical operational constraints are
- Communication windows with the ground control center during certain critical operations.
- Frequency of the ground updates, including the uncertainty and time delays of the upload parameters.
- Visibility constraints of the sensors during the reference trajectory, for instance operational range and illumination conditions.
- Sequencing of operations, for instance the acquisition time for sensors, the duration of slew maneuvers, and data transmission to ground.

From these top-level requirements, the engineering process will define the GNC architecture and will derive requirements for each component of the GNC system. Then, analysis and assessment of different options for each component will permit the selection of the optimal algorithms and equipment to fulfill all the requirements.

According to the validation and verification (V&V) plan, different tests will be performed following the selected software life cycle (e.g. V-cycle, spiral); see [Chap. 16](#). A typical sequence is unit, integration, system, and acceptance tests. The V&V of most of the requirements will require executing simulations (e.g. Monte Carlo, worst-case). During the

incremental stages in the development of a GNC system from low TRL to flight, the fidelity of the simulation environment increases accordingly. In order to accelerate the technology development process, there is a current trend to embed model-based development languages into real-time systems. The technology development plan would consist of a series of systems of progressively increasing TRL: model-in-the-loop (MIL), software-in-the-loop (SIL), processor-in-the-loop (PIL), and finally hardware-in-the-loop (HIL).

12.6.2 Orbit Navigation

The navigation function is typically implemented as a digital filter in the on-board computer. This must provide the necessary parameters to the guidance and control functions, for example current vehicle state. If just one sensor is able to provide all of the required information by the guidance and control functions, then the navigation algorithm might be a simple low-pass filter, for example a LIDAR system providing range, line of sight (LOS) and relative attitude during the last phases of rendezvous.

An example of a simple kinematic filter is the fading memory filter. This is presented in [Sect. 12.6.5](#) to filter the LOS provided by the camera and image processing during the approach to a point-source object. The second-order filter is formulated as follows (see [8] for the formulation of different orders)

$$\begin{aligned}\boldsymbol{\rho}_{j+1} &= \mathbf{y}_{j+1} - \left[\hat{\mathbf{y}}_j + (t_{j+1} - t_j) \dot{\hat{\mathbf{y}}}_j \right] \\ \dot{\hat{\mathbf{y}}}_{j+1} &= \dot{\hat{\mathbf{y}}}_j + (t_{j+1} - t_j) \ddot{\hat{\mathbf{y}}}_j + \left(1 - (1 - \sqrt{G})^2 \right) \boldsymbol{\rho}_{j+1} \\ \ddot{\hat{\mathbf{y}}}_{j+1} &= \ddot{\hat{\mathbf{y}}}_j + \left[G / (t_{j+1} - t_j) \right] \boldsymbol{\rho}_{j+1}\end{aligned}\tag{12.148}$$

where $\hat{\mathbf{y}}_j$ is the filter output (predicted value) at time t_j , \mathbf{y}_j is the input (measured value), and $\boldsymbol{\rho}_{j+1}$ is the *a priori* residual at time t_{j+1} . The second-order fading memory filter assumes a linear trend in the measurement (see the *a priori* residual formula). This filter requires only one configuration parameter, the gain G , which is a constant between zero and unity and is related to the memory length of the filter. Decreasing G makes the filter remember more previous measurements at the expense of decreasing the reactivity of the filter (i.e. decreasing the bandwidth).

In most situations, simple digital signal processing filters are not suitable because the navigation filter must estimate uncertain parameters not directly observed by the sensors (e.g. thrust level delivered by the propulsion system), and/or process measurements from different sensors (data fusion),

at different frequencies, and/or estimate the vehicle state during intervals without measurements (sensor black-out).

To achieve these navigation objectives, an optimal estimator is needed (e.g. a Kalman filter). A good example of the design of a navigation filter is presented in [12] for making a precise landing on Mars. The sensor suite includes a strapdown IMU, a phased-array radar and a scanning LIDAR. The radar and LIDAR provide altitude and surface-relative velocity data, with the radar being the primary sensor. The proposed state vector contains 18 parameters for estimated inertial position, velocity, attitude, and additional parameters that are related to sensor modeling (gyro bias, accelerometer bias, surface slope and altimeter bias).

The optimal state estimation filters can be classified as sequential or batch filters. Sequential filters process only the latest measurement while batch filters process a set of measurements taken during a certain time interval. Sequential filters respond more rapidly to unexpected variations in the state (e.g. detection of failures) but batch filters are more robust to modeling errors. Examples of batch filters are the weighted least-squares filter [9, pp 14–17, 10] and the Square-Root Information Filter (SRIF) [9]. The same notation given in Sect. 12.1.4 for the extended Kalman filter is followed in the next equations. The weighted least-squares filter aims at minimizing the weighted measurements residuals

$$J = (\mathbf{y}_j - \mathbf{H}_j \hat{\mathbf{x}}_j)^T \mathbf{W} (\mathbf{y}_j - \mathbf{H}_j \hat{\mathbf{x}}_j) \quad (12.149)$$

where \mathbf{y} is the vector of measurements, \mathbf{H} is the measurement sensitivity matrix, \mathbf{W} is the weight matrix, and $\hat{\mathbf{x}}$ is the estimate of the state. Minimizing the cost function J gives the optimal estimate of the uncertain parameters. The covariance of the estimation as follows

$$\begin{aligned} \hat{\mathbf{x}}_j &= (\mathbf{H}_j^T \mathbf{W} \mathbf{H}_j)^{-1} \mathbf{H}_j^T \mathbf{W} \mathbf{y}_j \\ \mathbf{P}_j &= (\mathbf{H}_j^T \mathbf{W} \mathbf{H}_j)^{-1} \mathbf{H}_j^T \mathbf{W} \mathbf{R}_j \mathbf{W} \mathbf{H}_j (\mathbf{H}_j^T \mathbf{W} \mathbf{H}_j)^{-1} \end{aligned} \quad (12.150)$$

where \mathbf{R}_j is the covariance matrix of the measurement noise.

The SRIF is derived from the least-squares filter. This formulation achieves higher numerical stability and precision at lower computational cost. These benefits become important when processing large batches of measurements. The batch filters can be used in sequential mode if the batch only includes the newest measurement. *A priori* information can be included as an additional measurement.

The most popular filter for on-board navigation is the Kalman filter [11], introduced in Sect. 12.1. It is a dynamic filter that is usually used in sequential estimation, although batch filtering is possible. There are several formulations

- The linear Kalman filter (LKF) or just Kalman filter that considers a linear time-invariant dynamical model (analytical transition matrix).

- The extended Kalman filter (EKF) that considers non-linear propagation of the average state.
- The extended Kalman but with U-D decomposition¹ of the covariance matrix to assure the conservation of the positive definiteness in the time and measurement updates.
- The unscented Kalman filter (UKF) in which the statistics are propagated and constructed from a set of wisely selected points (sigma points).

The formulation of the EKF and UKF were described in Sect. 12.1.4. The steps of the Kalman filter are schematically shown in Fig. 12.20. The time update propagates the average state and its covariance matrix from the last epoch (which may be the initialization epoch and state) to the current measurement time (a priori state and covariance). Then, a test of hypotheses is done on the input measurements. If the measurements are accepted, the measurement update provides the a posteriori state and covariance. The measurement update might be iterated in order to smooth the non-linearities. The a posteriori residuals and covariance are checked against the hypotheses and the convergence criterion. The output is prepared in the proper format and some information is made available for aiding the measurement management (e.g. for image processing).

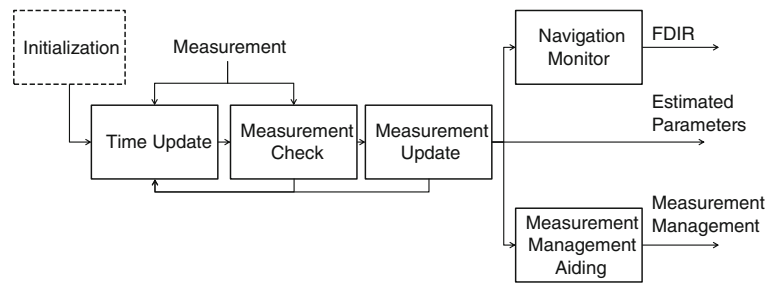
Different filters might be used in different phases of a mission. More usually, different configurations of a filter are required in different phases. The filter design must take into account the un-modeled physical effects and their impact in the propagation and observations, the uncertainty in the parameters of the considered dynamics and measurement models, the inclusion of multiple sensors in the navigation chain, and the allocated computer resources.

The augmented state vector refers to the uncertain parameters considered in the filter. The selection of its components must consider the above mentioned issues. It typically includes the spacecraft's state (in proper coordinates), and the uncertain parameters from the dynamics and measurements model.

The dynamics equation for each uncertain parameter depends on the parameter itself and the application. The most usual models for uncertain parameters from dynamics and measurements are biases (constant average value), drift (linear time dependency) or colored noises (e.g. exponentially correlated random variables [9]). Additive white

¹ The U-D decomposition avoids a problem of numerical stability (round-off error) in Kalman filters when the process noise covariance is small that can lead to a small positive eigenvalue being wrongly computed as a negative, causing the state covariance matrix to be indefinite when it should be positive-definite. The U-D decomposition, $\mathbf{P} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}^T$, where \mathbf{U} is a unit triangular matrix (with unit diagonal), and \mathbf{D} is a diagonal matrix, avoids some of the square root operations required by alternative methods, while maintaining their desirable properties.

Fig. 12.20 Functional architecture of a Kalman filter for autonomous navigation



Gaussian noise is usually included as process and measurement noise. These noises come from unknown or unmodeled sources. The process noise increases the a priori covariance during the time update. The measurement noise reduces the weight of the measurements in the measurements update. A high value of the variances of the noises, as provided to the filter, to avoid divergence of is sometimes called artificial noise (which might include process noise in the position differential equation).

12.6.2.1 Measurement Types

The measurements are the other fundamental element of the navigation function. The observation type and quality depends on the sensor and on the processing algorithm of its raw data. The performance of a navigation filter is strongly dependent on the type of measurements, their frequency, and their quality. The performance of other systems (e.g. pointing error and stability) affects the performance of the on-board sensors.

Often a single sensor cannot provide all of the required observables during all the operational ranges of a mission, so a sensor suite is mounted. In each phase, different sensors are used, or the same sensor with different processing algorithms (e.g. camera providing LOS at far distances, or providing LOS and range at close distances). The filter processes different measurement types to estimate the uncertain parameters (data or sensor fusion). An example is the hybridization of inertial navigation by an IMU with other sensor such as GPS or an altimeter. For the selection of the sensor suite, different issues must be taken into consideration. For instance, for far-range operations the nominal sensor might combine a large FOV with a coarse accuracy in order to assure observation in the presence of large position errors. However, for close operations, the nominal sensor is switched to another one with a narrower FOV and a higher accuracy in order to achieve better navigation performance. The following considerations must be traded in order to optimize the sensor set.

- What observables are needed for navigation.
- Which sensors provide the observables (note that some sensors can provide several observables simultaneously).
- What accuracy is required (accuracy should be one order of magnitude better than the overall GNC requirement,

but if this is not feasible, then at least two or three times better).

- What is the operational range (distance, FOV, velocity, angular rate ...).
- What are the system-level implications.
 - System constraints such as power, size, mass, cost.
 - Operational constraints such as illumination, on-board versus ground-based processing.

There are several sensors that might provide the same observable in a given mission phase. The combination of the different observables must provide the information required by the navigation filter (the system must be observable). Note that not all of the estimated parameters need to be directly observed. For instance, the velocity can be estimated by a dynamic filter after a certain time using only position-related observables (LOS and range). A list of different sensors that can provide the most common observables is given below. As technology evolves, new sensors appear that are more accurate, for instance the flash LIDAR or 3D time-of-flight (TOF), camera. Thus, the list below is not exhaustive but includes the most frequently used sensors for autonomous GNC applications. More details of specific sensors for rendezvous and formation flying will be given in Sect. 12.6.5.

- Range (distance) measurements are directly provided by altimeters (e.g. radar type in ESA's Huygens), range-finders (e.g. laser type in NASA's NEAR-Shoemaker), or RF sensors (e.g. in CNES's Formation Flying Radio Frequency system on the PRISMA mission). These are active sensors because they emit energy and record the returned signal. A summary of characteristics of altimeters and range-finders is provided in Table 12.6. Range can be derived from camera images (passive sensor) in certain applications by means of proper image processing (e.g. in ESA's ATV).
- Range-rate can be provided by RF sensors that measure the Doppler shift (e.g. the Russian Kurs system on Soyuz and Progress spacecraft), or by Doppler radar (e.g. the Viking landers).
- Line-of-sight can be provided by optical cameras (as in the approach phase of JAXA's Hayabusa mission), RF sensors (the Kurs system for rendezvous), or by imaging LIDAR sensors (like the scanning LIDAR in ATV that

Table 12.6 Characteristics of altimeters and range-finders for space navigation

	NLR [13]	Hayabusa LRF [14]	Hayabusa LIDAR [14]	MRA 1 [15]	MRA 2 [16]	Huygens [17]
Type	Laser range-finder	Laser range-finder	Laser range-finder	Radar altimeter	Radar altimeter	Radar altimeter
Heritage	NEAR	Hayabusa	Hayabusa	Beagle	–	Huygens
Manufacturer	Johns Hopkins Univ. (US)	NEC Toshiba (JP)	NEC Toshiba (JP)	Roke Manor (UK)	Roke Manor (UK)	Ylinen (FI)
Operational distance	<50 km (nominal; start to provide data @ 250 km)	LRF-S1: 7–120 m (4 heads) LRF-S2: 0.5–1.5 m	50 m to 50 km	1.5–700 m	0.2–100 m	100 m to 20 km
Range fccuracy (m)	6 m	LRF-S1: 0.1 m @ 10 m 3 m @ 100 m LRF-S2: 0.01 m	10 m @ 50 km 1 m @ 50 m	Normal: 0.5 m (1.5–700 m) High: 0.125 m (1.5–100 m) Low: 5 m (1.5–700 m)	0.02 m	2.6 m
Wavelength/frequency	1,064 nm	–	1,064 nm	4.2–4.4 GHz	76–77 GHz	15.4 and 15.8 GHz
Mass (kg)	5	2.16	3.6	0.4	0.4	1
Power (W)	16.5	8.6	22	3	3	8

provides simultaneously range and LOS). 3D ranging sensors based on either triangulation or time-of-flight (LIDAR) can provide relative position and orientation during close rendezvous or formation flying.

- Horizontal velocity (normal to the LOS) can be measured by a Doppler radar (as in NASA's Mars Science Laboratory) or by processing series of camera images (like DIMES in NASA's MER).
- Complete position information is provided by sensors such as space-qualified GPS (or GNSS) receivers that provide a PVT solution for a spacecraft in the vicinity of the Earth (as in JAXA's HTV), an Inertial Navigation System (INS) based on IMU measurements (as in the Ariane 5 or Vega launchers), or a 3D imaging sensor (like scanning LIDAR). The performance of some scanning LIDARs for the rendezvous application is presented in Table 12.7. INS provides position and velocity by integration of the equation of dynamics using the combination of gyro and accelerometer measurements of an IMU. Vectorial acceleration in the instrument frame determined by the accelerometers needs the attitude reference provided by the gyros to calculate the acceleration in the inertial frame. The performance of the INS navigation solution depends on the accuracy of the initial conditions, the gravitational model, and the non-gravitational acceleration measurements provided by the accelerometers. The accuracy of the INS solution degrades with time, and sometimes hybridization with other sensors is needed. For instance vision-based measurements or altimeters are envisaged for pinpoint landing on the Moon (several

hundred meters of landing dispersion). Hybridization of an IMU with other sensor might allow the use of lower IMU class (cheaper) and/or extend the duration without navigation performance degradation (multiple ignitions during launch).

Vision-based navigation is often used in many autonomous GNC systems. The main reasons are the low cost and system requirements (mass, size, power) of the cameras. Some examples of navigation cameras used for interplanetary navigation are presented in Table 12.8. Image processing can provide accurate LOS and range measurements (i.e. the full state), and even relative attitude. When observing point source objects, the accuracy of the LOS measurement is a fraction of the pixel angular size, i.e. the FOV/number of pixels. Figure 12.21 shows different observables that can be obtained from images of the Moon as an exemplar of an extended object. These observation types were traded for the design of the backup optical navigation system for crewed missions [19]. The main drawback is the image processing algorithms required to derive the observables. Often, some aids are used in order to simplify the image processing algorithms (more details in Sect. 12.6.5).

12.6.3 Orbit Guidance

Based on the estimation from the navigation function and the goals defined by the MVM, the guidance function must compute some or all the following outputs

Table 12.7 Performance of 3D imaging sensors for rendezvous

Scanning LIDAR Performance	MDA/Optech Spaceborne Scanning LIDAR System (SSLS)	Jena Optronik RVS (Rendezvous and Docking Sensor)
FOV	20° × 20°	40° × 40°
Measurement Range	2 m to 3 km	1 m to 2 km
Range Accuracy	0.05 m (3σ) @ short range	Noise: 0.1 m (3σ) @ long range Bias: 0.5 m @ long range Noise: 0.01 m (3σ) @ short range Bias: 0.01 m @ short range
LOS Accuracy	0.2° (3σ) @ short range	Noise: 0.1° (3σ) max Bias: 0.1°
Average Power (W)	<75	35 (nominal) 70 (max)
Total Mass (kg)	<10	6.1 (optical head) 7.7 (electronic box)

- Maneuvers to achieve the required goal, either impulses at certain times or thrust profiles for finite-thrust maneuvers.
- The reference trajectory (position and velocity) for a certain time interval in the future.
- Additional ephemerides required by other subsystems based on the updated maneuver plan.

The nominal trajectory is defined during the mission analysis, consolidated before launch and updated during flight. The computation of the reference trajectory and maneuvers often requires complex optimization algorithms that must fulfill all the operational constraints. In some other cases, the same guidance algorithms implemented on-board are used on the ground to design the reference trajectory and maneuvers. During the flight, perturbations such as maneuver execution errors, navigational uncertainties, operational delays and additional constraints, and disturbance forces, produce deviations from the reference trajectory that the guidance and control functions must cancel at the expense of additional propellant mass.

The main hypothesis is that the deviations from the reference trajectory are small and can be corrected with small variations in the reference thrust profile. This assumption permits the use of perturbation methods to compensate deviations from the trajectory. The guidance methods vary if the maneuvers are impulsive (approximated by an instantaneous change in velocity) or finite thrust (the duration of the thrusting arcs has a non-negligible impact in the trajectory). In case of low-thrust maneuvers, it is important to note that changing the thrust level (throttleable or pulse-modulated thrusters) and/or the thrust duration is required to be fully controllable. In addition, during long thrusting arcs it might be necessary to allocate short ballistic arcs for navigation tasks.

If the maneuvers are applied in open-loop (no control function), the guidance dynamical model must be sufficiently accurate to achieve the desired goal. If the maneuvers are executed in closed-loop, the control function can compensate small unmodeled effects in the guidance function.

Analytical algorithms that provide the solution in closed-form are preferred for on-board implementation. These solutions are not always available to achieve the guidance objective within the allocated error budget. It is important to note that the reference trajectory generated by the guidance algorithm must fulfill the operational constraints. Thus, in many cases the guidance problem is formulated as a constrained optimization problem.

One of the most useful guidance methods for impulsive maneuvers is based on the differential guidance. It was originally introduced for interplanetary navigation [18] but can be applied to rendezvous, formation flying, orbit maintenance, or descent and landing on small bodies. The basic formulation is

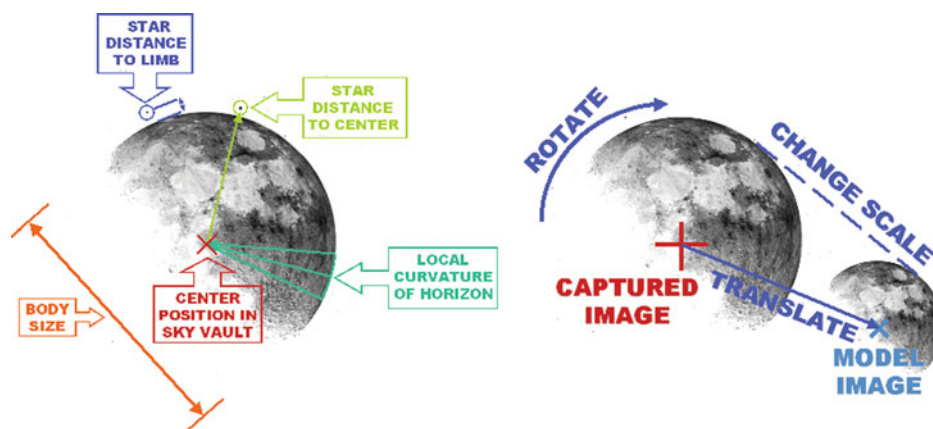
$$\begin{aligned} \begin{Bmatrix} \delta \mathbf{r}_N \\ \delta \mathbf{v}_N \end{Bmatrix} &\equiv \begin{Bmatrix} \mathbf{0} \\ \mathbf{0} \end{Bmatrix} = \Phi_{N,0} \begin{Bmatrix} \delta \mathbf{r}_0 \\ \delta \mathbf{v}_0 + \Delta \mathbf{V}_1 \end{Bmatrix} + \begin{Bmatrix} \mathbf{0} \\ \Delta \mathbf{V}_2 \end{Bmatrix} \\ \Phi_{N,0} &\equiv \frac{\partial \mathbf{x}_N}{\partial \mathbf{x}_0}. \end{aligned} \quad (12.151)$$

It considers an initial deviation from the reference trajectory ($\delta \mathbf{r}_0, \delta \mathbf{v}_0$) and two impulses, the initial delta-V ($\Delta \mathbf{V}_1$) which cancels the final position deviation at a fixed final time $\delta \mathbf{r}_N$, and the final delta-V ($\Delta \mathbf{V}_2$) which cancels the final velocity deviation $\delta \mathbf{v}_N$.

The linear system of equations defined by Eq. 12.150 can be solved explicitly. The key issue is the computation of the transition matrix $\Phi_{N,0}$. For linear time-invariant systems the transition matrix can be obtained analytically. In more complex dynamics, the transition matrix can be computed by

Table 12.8 Characteristics of narrow angle cameras used for navigation

	Miniature integrated camera and spectrometer (MICAS)	NavCam	Impactor targeting sensor (ITS)	Navigation camera (NC)	Framing camera (FC)	Optical navigation camera-telescopic (ONC-T)
Heritage	Deep space 1	Rosetta	Deep impact	Stardust	Dawn	Hayabusa
FOV (deg)	0.69×0.78	5×5	0.587×0.587	3.5×3.5	5.5×5.5	5.83×5.69
Detector array size (number of pixels)	$1,024 \times 1,024$	$1,024 \times 1,024$	$1,024 \times 1,024$	$1,024 \times 1,024$	$1,024 \times 1,024$	$1,024 \times 1,000$
Pixel size (μm)	9	13	21	12	14	12
Aperture (mm)	100	70	120	57.14	20	15
Focal length (mm)	677	152.5	2,100	202	150	120.8

Fig. 12.21 Illustration of star and limb related observables and image matching (positioning) from Moon. *Image* GMV

numerical differences or by integration of the variational equations of motion. Note that the second ΔV might never be applied since at the time of arrival at the final point a new maneuver can be calculated to achieve the next guidance objective (similar to a receding horizon control). The formulation can be extended to consider undefined final time. In this case, the final time is solved by minimizing the total ΔV .

The differential guidance can be applied at intermediate points of the trajectory to cancel perturbations that might otherwise grow to an unacceptable level at the final time (Fig. 12.22). These trajectory corrective maneuvers (TCM) can be applied at any instant if the navigation filter has converged after the previous maneuver.

When the effect of finite thrust on the trajectory is not negligible, impulsive guidance cannot be applied. If the thrusting duration is not too long compared with the guidance horizon, then a simple parameterization of the maneuver can be used. The small number of parameters permits the optimization of the maneuvers with low computational cost. An extreme case was implemented in Deep Space-1 [1]. This mission used electric propulsion to set up flybys of small bodies. Long low-thrust arcs were executed. The thrust profile in spherical coordinates was discretized as piece-wise linear expansion. A parameter optimization

problem was solved analytically. The gain matrix depended only on the sensitivity matrix of the final state deviation to the maneuver parameters.

When the thrust arcs are comparable to the guidance horizon (e.g. electric propulsion) more refined parameterizations are needed (for instance the guidance method in [21] was also applied to the descent and landing on small bodies). In some problems the optimal control theory [22, Chap. 5] can be applied. The parameters of the optimal control (initial co-states value) are solved to get the thrust profile that fulfills the guidance objective. This option can significantly increase the guidance computation time because the co-states (*aka* adjoints) are propagated simultaneously. There are exceptions when the adjoint dynamics can be solved analytically (e.g. optimal lunar descent and landing).

In some landing missions, hazard avoidance is required. The hazard avoidance system is usually not considered part of the GNC system. The measurements from the same sensors can be used by both, but the guidance will take as its objective the site selected by the hazard avoidance system. In turn, the hazard avoidance must consider the current navigation-estimated state and the guidance capabilities to compute the reachable site locus.

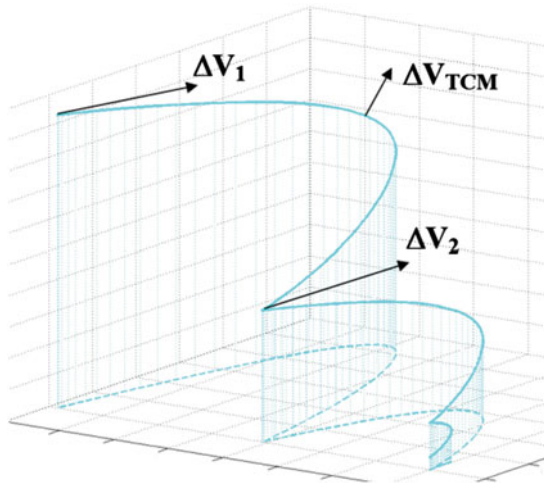


Fig. 12.22 Example of trajectory using differential guidance to achieve waypoints including intermediate corrective maneuvers. *Image GMV*

12.6.4 Orbit Control

Based on the current vehicle state estimation from the navigation filter and the reference trajectory computed at lower frequency by the guidance function, the main tasks of the translation control function are to cancel deviations that are not compensated by the guidance in closed-loop (controller proper), and to control the execution of the maneuvers that are computed by the guidance and the controller.

A closed-loop control system is presented in Fig. 12.19 (feed-forward and direct link are not depicted). Every component of the system introduces disturbances but these are not always additive (e.g. scale factors, cross-couplings). The closed-loop system must be stable in the presence of such disturbances. The disturbances from the sensors and actuators might be specified in the GNC design. In some cases these components are fixed and the GNC system must cope with their performances.

The output of the guidance function is a ΔV (impulsive maneuvers) or a thrust profile (finite thrust maneuvers). The control function must translate these maneuvers into the format required by the actuator management function. The most common actuators for translation motion control are thrusters. Other actuators can be solar panels or solar sails that modify the solar radiation acceleration vector, or aerodynamic control surfaces used in aero-assisted maneuvers (e.g. entry, descent and landing).

For typical thruster management algorithms, the translation control function must provide forces in spacecraft body axes. The desired force can be achieved by controlling either the thrust level during a given duty cycle (throttling), or the thrusting time with a constant thrust magnitude

(pulse-width in pulsed mode or total thrusting duration in steady mode). The thruster management function will select the optimal thrusters to provide the force, and also the torque requested by the attitude control function. The thruster management function will also define the thrusting time or thrust level of each thruster. Attitude constraints are considered in the guidance and controller algorithms. In practice, different thrusters might be used in different mission phases, for instance thrusters providing hundreds of newtons for large transfer maneuvers and newton-level thrusters for proximity operations.

The control function monitors the imparted acceleration in order to fulfill the guidance and controller commands. The control of the maneuver execution can be done in closed-loop or in open-loop. In open-loop the firing duration is computed from the expected acceleration. In closed-loop, state updates at high frequency are usually needed. The applied ΔV during one thruster control cycle can be measured by accelerometers if there are sensitive enough. If the acceleration provided by the thrusters is smaller than the resolution of the accelerometers, then other indirect measurements might be used. For instance, in NASA's Deep Space-1 the measurements of beam current and voltage were used to estimate the low thrust provided by the ion engine [20]. Comparison with thrust measurements made based on ground-based navigation showed a difference of less than 2 %.

In the case of a simple thruster architecture, the maneuver execution control can consider thruster saturation (maximum thrust) and minimum impulse bit (MIB), i.e. minimum the thrusting time, in the computation of the commanded ΔV . For finite thrust (usually low-thrust) a controller is usually implemented. The maneuver execution control is less demanding since the closed-loop controller frequently updates the thrust profile and can compensate for execution errors.

In some missions, the control function includes a closed-loop controller proper. This must cancel deviations produced by maneuver execution errors, dynamical perturbations, and uncertainty in the state. Thus, a controller increases the accuracy and robustness of the GNC at the expense of increased complexity and usually higher propellant expenditure. The presence of a translation controller is required when there are tight final delivery requirements (e.g. terminal rendezvous or precise landing).

The basic objective of the closed-loop translation controller of a space vehicle is to achieve the required performance and stability with low sampling rates. The performance of the controller is expressed in terms of different metrics, often competing with each other, such as low steady-state error and transient response error, sufficiently fast response time, and low propellant consumption. The

controller is implemented in software in the on-board computer. Hence, a discrete control is required. Discretization of continuous-time systems (e.g. bilinear or Tustin transformation) is often valid, but better performance can be achieved by designing the GNC system directly in the discrete domain.

Different trajectory control algorithms can be applied to different mission phases. A common approach in controller design is linearization around the reference trajectory provided by the guidance function. Some controllers often used for translation motion control are regulators and terminal controllers [22].

A regulator seeks to maintain a reference condition (e.g. a fixed position or an orbit). A state feedback regulator compares the reference state (guidance) with the estimated state (navigation) and generates an acceleration to cancel the error signal. Several methods exist to compute the control gain and fulfill the performance and stability requirements. The on-board implementation only needs to select the gain corresponding to the current mission phase and GNC mode.

The proportional-integral-derivative (PID) controller described in Sect. 12.2.5 is one of the most popular control techniques for single-input single-output (SISO) systems. The control acceleration is defined in Eq. 12.122. In the orbit GNC application the error signal is related to position (it might be an angle though). The proportional term reacts to a current error and is related to the response time (how fast the controller compensates a certain error). The integral term cancels the steady-state error of a pure proportional control. The integral term wind-up often appears in GNC systems (thruster saturation). The derivative term provides some prediction of the dynamics. It can help to decrease the settling time, to damp future oscillations and/or overshoots, and to increase stability (introduces a phase lead). If the derivative line is present, it is convenient that the navigation filter estimates the derivative of the state by minimizing the high frequency noise of the derivation.

The tuning of the controller consists of setting the different gains to achieve the requirements. There are several methods of tuning the PID controller (for instance the Ziegler-Nichols method) but they usually require manual trial and error. The controller often includes notch filters to avoid the excitation of lightly damped flexible modes. In addition, lag-compensation techniques are included when the delays introduced into the system overly reduce the stability margins.

The main advantage of PID control is that it does not require knowledge of the plant model (recall, the plant includes the actuators, the real world dynamics, and the sensors). However, it is only applicable when the channels can be decoupled. When cross-couplings prevent the use of

SISO control, modern multiple-input multiple-output (MIMO) methods are used. The optimal space-state methods require knowledge of the plant model in order to minimize a cost function. One simple optimal control technique is the linear-quadratic regulator (LQR), which assumes a linear (time-invariant) dynamics and a quadratic cost function J to compute the feedback control gain \mathbf{K} that minimizes J . The formulation for continuous control was presented in Sect. 12.1.5.

The weight matrices \mathbf{Q} and \mathbf{R} are the controller tuning parameters (the cross-coupling of control and state in the cost is usually not included). There are many software packages to obtain the optimal control gain \mathbf{K} for a given linearized dynamics (i.e. \mathbf{A} , \mathbf{B}). If there are different reference states, then a set of gains are pre-computed and stored on-board.

When there are uncertainties in the plant parameters, robust control methods are more convenient. These methods are mathematically cumbersome and require some knowledge of the plant. On the other hand, they can provide graceful degradation of performance in the presence of bounded uncertainties (if the deviation of the uncertain parameter from its nominal value is excessive then the system becomes unstable).

The non-linearities such as thruster saturation can introduce problems for stability and require proper treatment. In addition, dead-bands can be introduced to reduce propellant consumption. Finally, it is important to highlight the coupling between the navigation and guidance algorithms on the one hand, and the sensor and actuator equipment on the other. For instance, the navigation filter estimates the parameters requested by the guidance and control functions, removing the high-frequency noise from the sensors. Then, the guidance or control functions include the estimated biases (zero-frequency terms) in a feed-forward action in order to improve the GNC performance in terms of accuracy and propellant consumption.

A terminal controller seeks to achieve the desired conditions at a terminal time. An example is finite-horizon optimal control a problem of model predictive control that has been proposed for a variety of applications (pinpoint landing on Mars [22], low-thrust interplanetary trajectory control [21], and precise landing on asteroids [24]).

The trajectory is divided into N segments, in each of which the gravity field is approximated by a linear expansion at selected nodes \mathbf{r}_K . Considering a zero-order hold approach for the control acceleration and neglecting the other forces acting on the vehicle (e.g. no atmospheric drag), the resulting dynamics is a piecewise linear time-invariant (LTI) system. The equations of motion at a segment K are

$$\begin{aligned}\dot{\mathbf{x}}_K &= \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{G}_K & \mathbf{0} \end{bmatrix}}_{\mathbf{A}_K} \cdot \mathbf{x}_K + \underbrace{\begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}}_B \cdot (\tilde{\mathbf{g}}_K + \mathbf{u}_K) \\ \mathbf{G}_K &= \nabla \mathbf{g}|_{\mathbf{r}_K} = \frac{\mu}{r_K^3} \left(\frac{3}{r_K^2} \mathbf{r}_K \cdot \mathbf{r}_K^T - \mathbf{I} \right) \\ \tilde{\mathbf{g}}_K &= \mathbf{g}_K - \mathbf{G}_K \mathbf{r}_K \\ \mathbf{u}_K &= \mathbf{T}_K / m_K \\ \dot{m}_K &= -\frac{\|\mathbf{u}_K\|}{I_{SP} g_0} m_K\end{aligned}\quad (12.152)$$

where \mathbf{r}_K is the reference position at time t_K , \mathbf{x}_K is the state vector relative to \mathbf{r}_K , μ is the gravity parameter, \mathbf{I} is the 3×3 identity matrix, \mathbf{g}_K is the gravitational acceleration at t_K , \mathbf{T}_K is the reference thrust vector at t_K , m_K is the spacecraft mass at t_K , and I_{SP} is the specific impulse.

The constraints are to achieve the desired final state \mathbf{x}_f and to maintain the required thrust within the capability of the selected thruster (formulated as limited control acceleration a_{Kmax}). The cost function J is the sum of the squares of the control corrections $\delta \mathbf{u}_k$, in order to minimize the deviations from the optimal thrust profile (a quadratic objective function is convenient in solving guidance and control algorithms). Thus, the formulation of the discrete finite-horizon optimal control problem is

$$\begin{aligned}\min_{\delta\{\mathbf{u}_k\}} J &= \frac{1}{2} \sum_{k=0}^{N-1} \|\delta \mathbf{u}_k\|^2 \\ \text{subject to} & \\ \begin{cases} \mathbf{x}_k = \mathbf{A}_k \mathbf{x}_k + \mathbf{B}(\mathbf{u}_k + \delta \mathbf{u}_k) + \mathbf{B} \tilde{\mathbf{g}}_k, & \forall k = 0, \dots, N-1 \\ \mathbf{x}_N = \mathbf{x}_f \\ \|\mathbf{u}_k + \delta \mathbf{u}_k\| \leq a_{Kmax}, & \forall k = 0, \dots, N-1, \\ \text{where } a_{Kmax} = T_{max}/m_K, & m_K = m_{K-1} e^{-\|\mathbf{u}_{K-1} + \delta \mathbf{u}_{K-1}\| (t_K - t_{K-1}) / I_{SP} g_0}. \end{cases}\end{aligned}\quad (12.153)$$

These problems admit analytical transition matrix and closed-form solutions for the linear constraints. However, they need to be iterated in order to fulfill the non-linear constraints. Apart from the maximum thrust, attitude constraints are also non-linear and so must be considered [1, 21]. Nevertheless, the short computation time permits on-board implementation at a frequency sufficiently high to track the reference trajectory. A receding horizon implementation conveniently avoids singularities close to the terminal time.

12.6.5 Applications of Autonomous GNC

12.6.5.1 Rendezvous and Docking

The rendezvous phase consists of a series of operations that bring a chaser spacecraft from some thousands kilometers

to station-keep with a target spacecraft as a preliminary to mating with it. The convention that will be used in this section is that the origin of the relative coordinate system is located in the passive target and the chaser is the controlled spacecraft. The GNC system of the chaser is in charge of controlling the spacecraft state parameters in order to fulfill the mission and system constraints and to achieve the required docking or capture requirements. From the point of view of the GNC, berthing can be seen as a particular case of docking when the final relative velocity and angular rates are zero.

The rendezvous operation has several subphases that involve different GNC modes and equipment configurations. An example of autonomous rendezvous would be for a Mars sample return mission or an ATV approaching the ISS. The boundaries of the subphases depend on the sensors

- The initial phase is the *launch and orbit injection*. The chaser can be launched towards an orbiting target (e.g. an ATV launched towards the ISS), or the target can be launched towards an already orbiting chaser (e.g. a Mars Ascent Vehicle with the sample canister inside). The launch window must take into account the differential perturbations on the orbits of the target and chaser, the available on-board delta-V capability, and the time required for spacecraft activation, including the initialization of the relative navigation. The injection orbit must be designed to be passively safe, to require a total ΔV for the transfer to the target orbit that is within the available spacecraft budget, and to permit appropriate visibility windows for relative navigation acquisition. For instance, small launch time errors can result in expensive orbit plane corrections but can be naturally corrected by the J_2 differential node drift.
- The next phase is usually called *phasing or synchronization*. The main objective is to bring the chaser to a state or entry gate suitable for initiating the last stages of the rendezvous phase. Most of the delta-V budget of the rendezvous is expended in this phase, because it represents the largest changes in the orbit of the chaser. The definition of the synchronization strategy depends strongly on the navigation performance during this phase, on the timing to arrive at the entry gate, and on the maneuver execution errors.
- The *far-range rendezvous or homing* phase, or *intermediate rendezvous*, is where proper rendezvous operations start. It is defined by continuous visibility of the target with the far-range sensors, approximation of the relative motion (e.g. by Hill's equations), and the small maneuvers that are executed to approach to the target. This far rendezvous ends when short-range sensors can be acquired and there might be safety constraints on the location of the transition point (e.g. for the ISS the end of

far rendezvous must be outside a safety ellipsoid of $2 \times 1 \times 1$ km).

- The *close-range rendezvous* might be divided into the *closing and terminal rendezvous* subphases. The interface between these subphases is the safety corridor boundary. The closing phase seeks to acquire the entry corridor conditions with dispersions that are much less than the safety corridor dimensions. The safety constraints in this phase are as tight as they can be, and collision avoidance maneuvers will be executed in off-nominal conditions. The terminal rendezvous is the final phase, and usually involves a forced motion along a straight line with step-wise constant-approach rates. Continuous closed-loop control of the trajectory and attitude assures the achievement of the required final conditions for mating. The terminal rendezvous (sometimes known as the final approach) might end with a free drift to cover the last few meters (due to thruster efflux contamination issues or simply the availability of sensors).

During the design of the rendezvous trajectory, some ‘time-flexible elements’ [5] must be included in order to synchronize the rendezvous timeline with external events (visibility or communication windows) or schedules (crew or ground operations). The rendezvous timeline will need to be modified in-flight due to the uncertainties and errors/deviations that appear in real-world operations. Typical time-flexible elements are hold points (i.e. constant relative position) and free drift orbits.

The rendezvous problem has been traditionally analyzed in a circular orbit. However, elliptic rendezvous is nowadays often considered, if not for nominal operations (elliptic orbits might yield better mission performance) at least for contingency scenarios (e.g. orbit injection error). For the most useful maneuvers in circular rendezvous [5], equivalent maneuvers in elliptic rendezvous have been defined [23], including time-flexible elements (with the hold points being periodic orbits around the hold point). In addition, the analytical transition matrix of the Hill-Clohessy-Wiltshire equations (see Chap. 4) for circular orbits has an analogous closed-form solution for the elliptic rendezvous [39].

In order to estimate the relative position of the chaser with respect to the target, a sensor suite needs to be selected. Different sensors are available for relative navigation in different scenarios. A trade-off for each particular mission is needed. Table 12.9 summarizes the current sensors available for rendezvous. Note that the orbit of the target or the chaser must be known in order to formulate the equations of the relative motion. The reference orbit can be known from the ground-based orbit determination system or from satellite navigation systems like GPS (only in Earth orbit).

The relative state can be obtained from differences of absolute measurements (if available from both vehicles and known by the chaser GNC) but can lead to larger errors than direct observations of the relative state.

Providing specific values for sensor selection is difficult, because many parameters are mission dependent and must be assessed specifically. For instance, the size of the target has a significant effect on the operational range of some sensors (optical camera, RF and LIDAR on a target that does not have reflectors). However, some guidelines for sensor assessment can be provided for a rendezvous GNC system.

- RF-sensors are good for medium to long distances (mainly homing and closing) including acquisition and contingency (omni-directional or scanning antennas).
- GPS is a reliable well-known technology but requires a cooperative target and is only available operationally in Earth orbit. Relative kinematic GPS (using phase) provides sufficient performance (several cm with multipath) for use in closing and terminal rendezvous.
- Cameras with different fields of view can cover the entire rendezvous phase. The main problem for target acquisition and orbit synchronization is the visibility constraints (illumination conditions, faint target). The visibility windows for a narrow-angle camera and for an RF-sensor (omni-directional antennas on the target) are shown for a Mars rendezvous scenario in Fig. 12.23.
- Imaging LIDAR is a mature, robust, precise sensor that can provide relative position and attitude measurements at relatively high frequency (~ 10 Hz). Its main drawbacks are the limited operational range (only for close rendezvous) and the high cost, mass, and power consumption compared to other relative sensors.

Vision-based navigation is one of the most promising technologies for autonomous rendezvous, in particular when the target is non-cooperative one that does not incorporate aids for relative navigation. The cameras are often used in conjunction with other sensors in order to increase the robustness (for instance omni-directional RF-sensors for a higher localization probability in the case of large uncertainties or LIDAR for the terminal phase). For redundancy, different optical heads and processing units are mounted. It is convenient that the MVM can configure any combination of optical head and processing unit.

Different cameras might be necessary to cover the entire rendezvous. For instance, for optical far-range, a camera with narrow or moderate field of view (higher resolution) is best. During the far imaging, the object appears as a point source, and only the line of sight can be computed. The maximum range depends on the sensitivity of the camera,

Table 12.9 Sensors for relative navigation in rendezvous of spacecraft

Sensor	Measurements	Comments
Satellite navigation (e.g. GPS)	<ul style="list-style-type: none"> • Relative position • Absolute position • Relative attitude (several antennas) 	<ul style="list-style-type: none"> • Earth orbits only • Cooperative targets (GPS receiver in both spacecraft and communication link) • Maximum operational range limited by communication link (Earth shadowing) • Minimum range limited by shadowing and multipath • Reference orbit w/o ground intervention (absolute positioning and dynamic filter) • Two possibilities: absolute position subtraction or relative-GPS (raw measurements jointly processed, provides better accuracy) • Relative GPS using pseudo-range or carrier phase (much higher accuracy but ambiguity resolution needs longer initialization) • Coarse relative attitude due to short baselines and multipath • Low mass, power and cost (inter-satellite communication equipment not included) • In-flight heritage for RDV (e.g. ATV [5], PRISMA [41])
Radio frequency (RF)	<ul style="list-style-type: none"> • Range • Range-rate • LOS • Relative attitude (at least two receiving antennas) 	<ul style="list-style-type: none"> • Uncooperative (no equipment on target) or cooperative targets (transmitter, retro-reflectors on target) • Few visibility constraints: mounting (occultation, multipath, interference) • Wide operational range: ~ 10 m to ~ 100 km (same sensor would provide coarse resolution) • LOS measurement: fine (two antennas and carrier phase) or coarse • Similar algorithms than relative GPS. • Moderate/high mass, power and cost (operational range dependent) • Long heritage in crewed missions (e.g. Gemini, Soyuz/Progress [5])
Optical camera	<ul style="list-style-type: none"> • LOS • Range (pattern recognition, image matching, photometry) • Relative attitude (pose estimation with patterns or shape model) 	<ul style="list-style-type: none"> • Cooperative (patterns, LEDs) or uncooperative targets (shape known and complex image processing) • Several operational constraints: <ul style="list-style-type: none"> – Illumination conditions (geometry Sun-target-chaser), flash (illuminator) or LEDs on target to increase the visibility windows, e.g. during eclipses – Exclusion angle with bright objects (e.g. Sun, Earth and Moon) to avoid blinding, dazzling, blooming (APS detectors can relax constraints on exclusion angles) – Stray-light, reflections on chaser surfaces – Shadowing by chaser mechanisms (e.g. capture mechanism) • Specific image processing algorithms required to obtain measurements (might be cumbersome for uncooperative targets) • Wide operational range: ~ 10 m to ~ 100 km (with different FOV and image processing algorithms) • Low mass, power and cost • Short in-flight heritage, demonstration missions (e.g. ATV [5], PRISMA [38])
Imaging LIDAR	<ul style="list-style-type: none"> • Range • LOS • Relative attitude (at least three markers) 	<ul style="list-style-type: none"> • Retro-reflectors on target for better performances (image processing techniques for uncooperative targets) • Short operational range • High mass, power and cost • In-flight heritage for RDV (e.g. ATV [5]) • Flash LIDAR still not space-qualified and valid only for short ranges

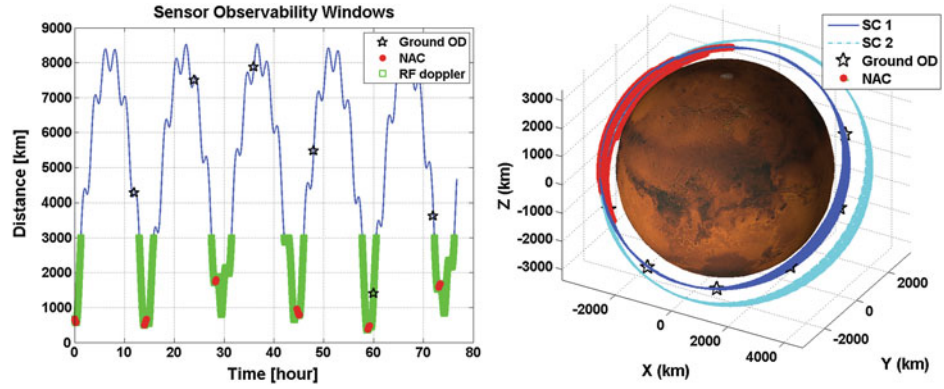
usually defined by the limiting magnitude. The image integration time is a critical parameter for far-range target detection. Long exposure times permit detection of fainter objects. However, a short exposure time is preferable to relax the attitude stability requirements. Some techniques permit these competing requirements to be traded off.

In optical close-range operations, the object is extended and a camera with larger field of view is preferred. In this phase, the shape of the object is distinguishable with sufficient resolution for the image processing function to

provide a distance to the target and even a relative attitude. Different image processing techniques are available to provide the required observables.

As discussed in Chap. 4, Hill's equations, also called the Clohessy–Wiltshire equations of motion, are a system of linear equations that approximate the relative motion between two bodies in orbit. These equations have been widely used since the early space missions in order to compute the dynamics of vehicles in a rendezvous scenario. Two hypotheses are assumed in deriving the general equations of

Fig. 12.23 Visibility arcs with nominal narrow-angle camera constraints (16° exclusion angle, 11th magnitude limit, eclipses). *Image GMV*



motion, a close distance between the two spacecraft, and circular orbits. Hill's equations are expressed as

$$\begin{aligned} \ddot{x} - 2\omega\dot{z} &= F_X/m \\ \ddot{y} + \omega^2 y &= F_Y/m \\ \ddot{z} + 2\omega\dot{x} - 3\omega^2 z &= F_Z/m \end{aligned} \quad (12.154)$$

where x, y, z are the chaser position with respect to the target in the local orbital frame of the target (Fig. 12.24), ω is the angular orbital rate of the target satellite, m is the mass of the chaser vehicle, and F_X, F_Y, F_Z are the differential forces acting on the chaser. Note that the reference frame used in deriving Eq. 12.153 differs from that used in the discussion of Hill's equations in Chap. 4, where z and x, y and z , and x and y are switched.

This system can be represented as a linear time-invariant system in the state space given below. Note that the out-of-plane motion (y) is decoupled from the in-plane motion (x, z). The range of validity of the Hill's equations is increased if the reference frame is formulated in curvilinear coordinates instead of Cartesian coordinates.

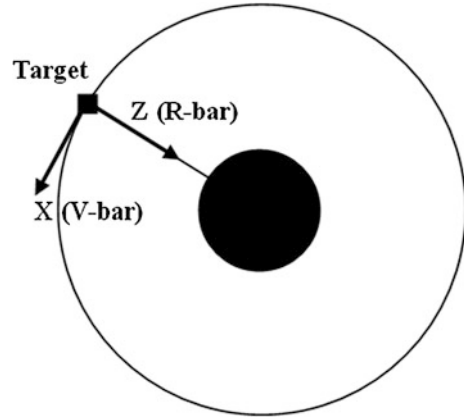


Fig. 12.24 Target local orbital frame

known as CW equations in the literature. The CW equations in the state space result in

$$\begin{Bmatrix} \mathbf{r}(t_F) \\ \mathbf{v}(t_F) \end{Bmatrix} = \begin{bmatrix} \Phi_{rr}(t_F, t_0) & \Phi_{rv}(t_F, t_0) \\ \Phi_{vr}(t_F, t_0) & \Phi_{vv}(t_F, t_0) \end{bmatrix} \begin{Bmatrix} \mathbf{r}(t_0) \\ \mathbf{v}(t_0) \end{Bmatrix} \quad (12.156)$$

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathbf{Ax} + \mathbf{Bu} \\ &= \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2\omega \\ 0 & \omega^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3\omega^2 & -2\omega & 0 & 0 \end{bmatrix} \begin{Bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{Bmatrix} \\ &+ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/m & 0 & 0 \\ 0 & 1/m & 0 \\ 0 & 0 & 1/m \end{bmatrix} \begin{Bmatrix} F_X \\ F_Y \\ F_Z \end{Bmatrix} \end{aligned} \quad (12.155)$$

The homogeneous solution (zero-input) of the Hill's equations (Eq. 12.153 or 12.154) is known as the Clohessy–Wiltshire (CW) equations and sometimes Eq. 12.154 is

where

$$\begin{aligned} \Phi_{rr}(t_F, t_0) &= \begin{bmatrix} 1 & 0 & 6(\omega\Delta t - \sin \omega\Delta t) \\ 0 & \cos \omega\Delta t & 0 \\ 0 & 0 & 4 - 3 \cos \omega\Delta t \end{bmatrix} \\ \Phi_{rv}(t_F, t_0) &= \begin{bmatrix} \frac{4}{\omega} \sin \omega\Delta t - 3\Delta t & 0 & \frac{2}{\omega} (1 - \cos \omega\Delta t) \\ 0 & \frac{\sin \omega\Delta t}{\omega} & 0 \\ \frac{2}{\omega} (\cos \omega\Delta t - 1) & 0 & \frac{\sin \omega\Delta t}{\omega} \end{bmatrix} \\ \Phi_{vr}(t_F, t_0) &= \begin{bmatrix} 0 & 0 & 6\omega(1 - \cos \omega\Delta t) \\ 0 & -\omega \sin \omega\Delta t & 0 \\ 0 & 0 & 3\omega \sin \omega\Delta t \end{bmatrix} \\ \Phi_{vv}(t_F, t_0) &= \begin{bmatrix} 4 \cos \omega\Delta t - 3 & 0 & 2 \sin \omega\Delta t \\ 0 & \cos \omega\Delta t & 0 \\ -2 \sin \omega\Delta t & 0 & \cos \omega\Delta t \end{bmatrix} \end{aligned} \quad (12.157)$$

$\Delta t = t_F - t_0.$

Fig. 12.25 Guidance and control of V-bar hopping (or radial impulse transfer along V-bar)

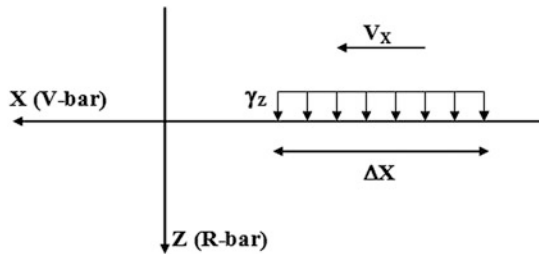
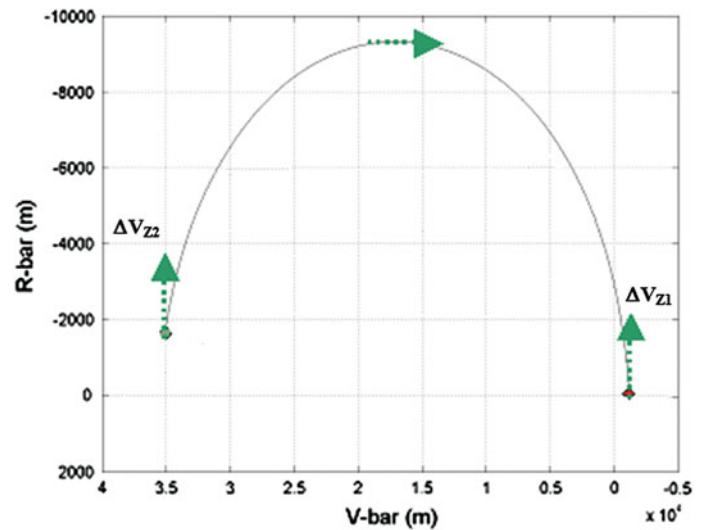


Fig. 12.26 Forced motion approach along the V-bar straight line

These equations are very convenient for use in a linear Kalman filter because the transition matrix is analytical. Hence, the time update (or propagation of the state and covariance matrix) is very fast. In addition, the analytical transition matrices are well suited for the differential guidance presented in Eq. 12.150. This discrete control can be applied in the case of impulsive transfers. Multiple rendezvous strategies for the circular case are given in [5, Chap. 3]. Some of them are extended for the elliptic rendezvous in [23]. For instance, in the case of V-bar hopping (Fig. 12.25), the nominal maneuvers for each hop are given by

$$\Delta V_{Z1} = \Delta V_{Z2} = \frac{\omega}{4} \Delta x \quad (12.158)$$

where Δx is the actual distance to be traversed. The duration of the hop maneuver is half an orbital period T . At an intermediate point, a correction maneuver can be applied using the transition matrix presented in Eq. 12.156.

In the case of forced motion along the V-bar (Fig. 12.26), the task of the guidance function is to compute the reference R-bar acceleration that will maintain a constant approach velocity V_x , considering the actual V-bar position from the

navigation. The approach velocity V_x is usually defined from safety considerations. The continuous R-bar acceleration γ_z is given below. The duration Δt depends on the approach velocity and the traversed distance Δx . Continuous control algorithms could be applicable in this continuous-thrust maneuver, with the considerations for discrete implementation mentioned in Sect. 12.6.4., namely

$$\begin{aligned} \gamma_z &= 2\omega V_x \\ \Delta t &= \frac{\Delta x}{V_x}. \end{aligned} \quad (12.159)$$

12.6.5.2 Terminal Phase of Missions to Small Bodies

When approaching an asteroid or comet, the spacecraft can take observations of the target with on-board sensors. The attainment of these relative measurements marks the start of the terminal phase of a mission to a small body. Thus, in this phase the spacecraft state relative to the target can be directly estimated (though sometimes not the full state).

The terminal phase poses different requirements and constraints for high-speed impact than for rendezvous missions. A flyby of a small body presents many similarities with an impact mission. For instance, NASA-JPL's AutoNav system [26] was used in Deep Space-1 for the flyby of comet Borrelly, in Stardust for the flyby of asteroid Annefrank, and in Deep Impact [40] for both the impactor and the flyby spacecraft.

An important factor is the communication delay, typically of tens of minutes, which is critical in hypervelocity impact or in descent and landing. Thus, autonomous operation in certain critical phases is mandatory. Due to the demanding delivery requirements, uncertain environment,

and tight operational constraints, the GNC system for the terminal phase is a critical enabling technology for missions to small bodies. For example, in an impact with a small body that is several hundred meters in size, the delivery accuracy must be several tens of meters.

Any mission to a small, irregular body has to cope with uncertain environmental conditions, namely the spacecraft's dynamical environment, and the shape, rotational state and surface characteristics of the target. In addition, the forces acting on the spacecraft are all small. Thus, the uncertainties in the dynamical model have a significant effect on trajectory prediction and in critical phases such as descent and landing.

The most generic methods to obtain relative observations are images with a camera (in the visual or near-infrared spectral range), and/or range measurements with a range-finder (radar or laser). The range output can be a digital elevation map (DEM) using a 3D sensor such as an imaging LIDAR. Typically, camera measurements can be taken at a greater distance than for a range-finder.

At the beginning of the terminal phase, the first objective of the GNC system is to detect and identify the small body against the starry background. Small, irregular objects observed at long distances ($\sim 10^6$ km) are point sources with low visibility. Thus the object is very faint and its brightness is highly variable. The observability depends on the rotational state of the object and the Sun-target-spacecraft relative geometry (distances and angles).

Long image integration times increase the signal-to-noise ratio. A tight relative pointing error (RPE) is required in order to concentrate the photons around a single pixel. During long integration times without tight RPE, a point source will produce a characteristic pattern in the image. For this imaging strategy, Deep Space-1 implemented an image processing technique called multiple cross-correlation [26]. Such image processing algorithms correlate each object in an image with a mask template extracted from the same image for center finding.

For target detection and point-source tracking, a narrow-angle camera ($\text{FOV} \lesssim 5^\circ$) is the best option because it provides higher sensitivity and line of sight accuracy. The line of sight accuracy is typically a fraction of the pixel angular size (or pixel $\text{FOV} = \text{FOV}/\text{No. pixels}$). A star tracker ($\text{FOV} \sim 10^\circ$) can be a good alternative in order to reduce the equipment carried if the reduction in accuracy and sensitivity is acceptable.

Early maneuvers are more effective for correcting deviations, but navigation accuracy is usually worse at larger distances. Hence, it is necessary to design in parallel the navigation chain, the guidance strategy, and the control system to achieve the delivery requirements with optimal use of the on-board resources.

In the case of an impact or flyby mission, the time-to-go (time to impact or to minimum distance) is not controlled. Therefore, only deflection maneuvers are required to control the impact point. Several guidance and control strategies can be applied [28], involving different propulsion systems.

- Predictive-impulsive guidance, where impulsive maneuvers are executed at predefined times. This strategy is suitable for relatively high thrust (e.g. tens of newtons for vehicles of several hundred kilograms).
- Proportional navigation, where continuous thrust is proportional to the line of sight rate and approach speed. This is a well-known method for missile guidance. It is particularly suitable for low-thrust propulsion (e.g. less than 1 N for vehicles of several 100 kg). The minimum thrust required to compensate for the initial deviation of the impact point must be considered.

- A hybrid scheme that implements mid-course predictive-impulsive guidance and terminal proportional-navigation. It is designed for missions with intermediate thrust levels.

The navigation filter is closely related to the impact guidance and control strategy. For the above mentioned strategies, two types can be implemented

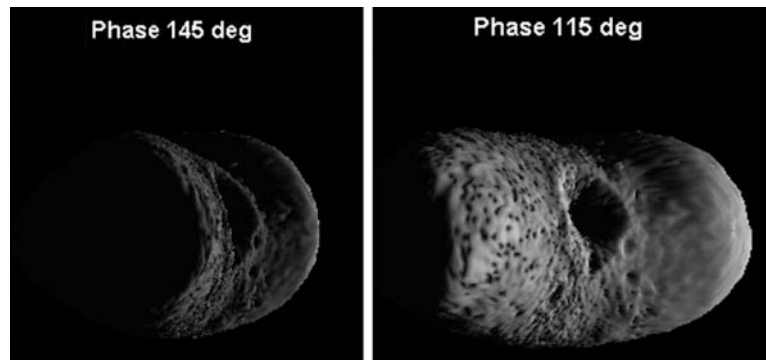
- Estimation of the line of sight and its time derivatives. Some applicable filters are digital fading-memory or batch-sequential least squares. These filters are more sensitive to the image processing performance.
- Estimate the complete relative state vector with a dynamic filter (e.g. a Kalman-Schmidt filter [29]). These filters are prone to overestimate the innovation (or measurement residual) vector and also suffer numerical problems due to the non-observability of some components of the state vector.

The rendezvous missions present notable differences. During the approach phase most of the relative velocity is canceled by means of braking maneuvers. The duration of this phase is long enough to involve ground control in the navigation chain. This relaxes the requirements in the image processing, although due to the communication limitations some critical pre-processing operations might still be done on-board. Thus, a fully autonomous GNC is not mandatory (different levels of autonomy might be used depending on the particular mission).

This approach can be split in two phases, far and close approach. During the *far approach*, the celestial object is a point source in the image. It starts after successful detection and identification of the small body. Asteroids of ~ 1 km are typically resolved (extend over several pixels) at a range of $\sim 1,000$ km. This marks the end of the far approach.

Laser range-finders used in past missions to small bodies achieved target acquisition at 250 km (NEAR-Shoemaker) and 50 km (Hayabusa). Photometry can be used to infer range from the brightness of the object. However, the

Fig. 12.27 Simulated images of an Eros-shaped asteroid at the point of peak brightness (minimum integrated magnitude); axis of rotation perpendicular to the approach plane. *Image GMV*



measurements are subject to very large error (up to 100 % depending on prior knowledge of the object's lightcurve and properties). Thus, during far approach, the navigation function must estimate the relative distance and speed without direct range measurements. The more accurate options are

- Long ballistic flight allowing long observation arcs (spanning a significant portion of the object's orbit). It takes a long time, and the relative trajectories must provide good observability conditions.
- Execute 'dog-leg' maneuvers in order to rapidly change the observation geometry [30]. Proper design of the approach trajectory (and thus the maneuvers) is mandatory.

It is important to note that due to the irregular shape of most small bodies there is a risk of losing tracking during the far approach (Fig. 12.27). The image processing and navigation filter must be robust against such an event [30]. The accuracy of the navigation filter depends on prior knowledge of certain characteristics of the object. The sequence of operations must assure that there is sufficient knowledge of the target to accomplish the next phase objectives.

In the *close approach* the target appears as an extended object in the camera frame. In addition, if a range-finder is mounted, range measurements become available within a certain distance. The image processing algorithm might be based on center of brightness (CoB) computation. Other more accurate and complex image processing techniques require knowledge of the size and shape of the object. The center of brightness–center of mass (CoB–CoM) offset should be considered in the navigation filter. The GNC algorithms can be applied in certain modes of the proximity operations.

In the *proximity operations*, the gravity of the target becomes non-negligible, and will be the most significant force during descent and landing. Still, the dynamics are slow and a fully autonomous GNC is not mandatory, except for descent and landing. The weak, irregular gravity field opens new possibilities for orbiting the target, like hovering (station-keeping) or self-stabilized terminator orbits (a.k.a. photo-gravitational orbits). The selection of the most

appropriate sequence of orbits for a certain mission depends mainly on the uncertainties of the gravity field (safety issues), surface observability (for science and navigation), and propellant expenditure.

When the target has a large angular size, a wide angle camera ($FOV \gtrsim 20^\circ$) is the best option for vision-based navigation. The image processing can use different techniques to obtain relative measurements. The Hayabusa mission demonstrated that the necessary image processing can be done on the ground. If the image processing is done autonomously, then a dedicated processing unit might be needed. If good knowledge of the size and shape of the asteroid is available, limb-related measurements (limb-star angular distance or star occultation time) or image matching can be obtained.

If the images have sufficient resolution of the surface of the target, then known landmark identification or unknown feature tracking is feasible. Known landmark mapping permits direct positioning of the spacecraft state in the target's body-fixed frame. Unknown landmark tracking provides measurements of the velocity relative to the surface. Thus, unknown landmark tracking must be used in combination with other measurements for complete state estimation.

An interesting option to simplify the image processing is to deliver markers to the surface, which was the Hayabusa strategy. The markers can either include LEDs or be illuminated using a flash, to make them the brightest objects in the image. The markers serve as beacons (artificial landmarks) to land in a given position nearby.

The use of an altimeter for navigation is mandatory if information on the shape and size of the target is not available from previous phases (from science observations or ground-based navigation). In any case, altimetry significantly increases the robustness and accuracy of the navigation system. If a 3D sensor (e.g. imaging LIDAR) can be used, then the 3D images can be used for terrain-relative navigation. Such 3D images provide more information than combined optical images and altimetry, but need to be traded with the total cost, mass, power and volume. Note

that redundancy in the navigation chain is needed for a single-failure tolerant system.

The objective of a *descent and landing* (D&L) might be either to land softly at a given point or to hover at a very low altitude above a selected site. The GNC must be fully autonomous below the so-called low gate, in analogy the Apollo with lunar missions. Some navigation sensors (e.g. optical cameras or imaging LIDAR) must avoid large angular rates and maintain the landing site continuously in view. Thus, for trajectory control the propulsion system should be able to provide pure force (no torque) in any direction, i.e. no rotation is required in order to apply force in any direction.

The orientation of the spacecraft with respect to the surface of the target must assure that there is no risk of collision of solar arrays or tumbling after touchdown. Thus, a 3D sensor or a number of tilted altimeters should provide information of the orientation of the vehicle with respect to the surface. The sensors must be assured of good visibility during all phases. For instance, a star tracker must not be blinded or dazzled during the proximity operations, and the legs must not appear in the wide-angle camera images.

References

- Wertz, J. R. (ed.), *Spacecraft Attitude Determination and Control*, Kluwer Academic Publishers, The Netherlands, 1978.
- Leondes, C. T. (ed.), *Guidance and Control of Aerospace Vehicles*, McGraw-Hill, 1963.
- Bryson, A.E., *Control of Spacecraft and Aircraft*, Princeton University Press, Princeton, NJ, 1994.
- Sidi, M. J., *Spacecraft Dynamics and Control: A Practical Engineering Approach*, Cambridge University Press, Cambridge, 1997.
- Wie, B., *Space Vehicle Dynamics and Control*, American Institute of Aeronautics and Astronautics, Second Edition, Restone, VA, 2008.
- Pisacane, V. L. (ed.), *Fundamentals of Space Systems*, Oxford University Press, 2nd Edition, 2005, Chapter 5.
- Fortescue, P., Swinerd, G., and Stark, J. (eds.), *Spacecraft Systems Engineering*, 4th Edition, John Wiley & Sons, 2011, Chapter 9.
- Wertz, J. R., Everett, D. F., and Puschell, J. J. (eds.), *Space Mission Engineering: The New SMAD*, Microcosm Press, Hawthorne, CA, 2011, Chapter 19.
- Wahba, G., "A Least-Squares Estimate of Satellite Attitude," *SIAM Review*, Vol. 7, No. 3, July 1965, p. 409.
- Black, H. D., "A Passive System for Determining the Attitude of a Satellite," *AIAA Journal*, Vol. 2, No. 7, 1964, pp. 1350-1351.
- Keat, J., "Analysis of Least Squares Attitude Determination Routine DOAOP," Computer Sciences Corp., CSC/TM-77/6034, Silver Spring, MD, Feb. 1977.
- Shuster, M. D. and Oh, S. D., "Three-Axis Attitude Determination from Vector Observations," *Journal of Guidance, Control, and Dynamics*, Vol. 4, No. 1, 1981, pp. 70-77.
- Markley, F. L., "Attitude Determination Using Vector Observations and the Singular Value Decomposition," *Journal of the Astronautical Sciences*, Vol. 36, No. 3, 1988, pp. 245-258.
- Shuster, M. D., "A Simple Kalman Filter and Smoother for Spacecraft Attitude," *Journal of the Astronautical Sciences*, Vol. 37, No. 1, 1989, pp. 89-106.
- Bar-Itzhack, I. Y., "REQUEST: A Recursive QUEST Algorithm for Sequential Attitude Determination," *Journal of Guidance, Control, and Dynamics*, Vol. 19, No. 5, 1996, pp. 1034-1038.
- Choukroun, D., Bar-Itzhack, I. Y., and Oshman, Y., "Optimal-REQUEST Algorithm for Attitude Determination," *Journal of Guidance, Control, and Dynamics*, Vol. 27, No. 3, 2004, pp. 418-425.
- Lefferts, E. G., Markley, F. L., and Shuster, M. D., "Kalman Filtering for Spacecraft Attitude Estimation," *Journal of Guidance, Control, and Dynamics*, Vol. 5, No. 5, 1982, pp. 417-429.
- Crassidis, J. L. and Markley, F. L., "Unscented Filtering for Spacecraft Attitude Estimation," *Journal of Guidance, Control, and Dynamics*, Vol. 26, No. 4, 2003, pp. 536-542.
- Crassidis, J. L. and Junkins, J. L., *Optimal Estimation of Dynamic Systems*, Chapman & Hall/CRC, Boca Raton, FL, 2004.
- Crassidis, J. L., Markley, F. L., and Cheng, Y., "Survey of Nonlinear Attitude Estimation Methods," *Journal of Guidance, Control, and Dynamics*, Vol. 30, No. 1, 2007, pp. 12-28.
- Bryson, A. E., *Applied Linear Optimal Control*, Cambridge University Press, 2002.
- Bryson, A. E. and Ho, Y-C, *Applied Optimal Control*, John Wiley & Sons, 1975.
- M.S. Hodgart "Gravity Gradient and Magnetorquing Attitude Control for Low Cost Low Earth Orbit Satellites- the UoSAT Experience", Ph.D. Submission at University of Surrey, June 1989
- Brady, T., Tillier, C., Brown, R., Jimenez, A., and Kourepenis, A., "The Inertial Stellar Compass: A New Direction in Spacecraft Attitude Determination," SSC02-II-1, 16th Annual USU Conference on Small Satellites, 2002.
- EADS Sodern Horizon Sensor, http://www.sodern.com/site/FO/scripts/siteFO_contenu.php?mode=&noeu_id=61&lang=EN, [Accessed 29 March 2011].
- SELEX Galileo Horizon Sensor, http://www.selex-sas.com/EN/Common/files/SELEX_Galileo/Products/IRES_NE.pdf, [Accessed 29 March 2011].
- Van Bezooijen, R.W.H, A Star Pattern Recognition Algorithm for Autonomous Attitude Determination, IFAC Automatic Control in Aerospace, Japan, 1989
- Steyn, WH, A Multi-mode Attitude Determination and Control System for Small Satellites, PhD Thesis, Stellenbosch, 1995
- Lightsey, G.E., "Spacecraft Attitude Control Using GPS Carrier Phase", *Global Positioning System: Theory and Application*, Parkinson, B.W. and Spilker, J.J. (ed), Vol. II, 1996.
- Lightsey, G.E., Cohen, C.E., Feess, W.A. and Parkinson, B.W., "Analysis of Spacecraft Attitude Measurement Using On-board GPS", *Advances in the Astronautical Sciences*, Vol. 86, 1994.
- Lappas, Vaios J (2002) A Control Moment Gyro (CMG) Based Attitude Control System (ACS) For Agile Small Satellites Doctoral thesis, University of Surrey.
- Honeywell M50 Data Sheet, http://www51.honeywell.com/aero/common/documents/myaerospacecatalog-documents/M50_Control_Moment_Gyroscope.pdf, Accessed 23 March 2011
- Burt, Richard "AAS 03-072 "Failure analysis of International Space Station Control Moment Gyro" 26th Annual AAS Guidance and Control Conference, Breckenridge, Colorado.

34. http://www51.honeywell.com/aero/common/documents/myaero-spacecatalog-documents/M50_Control_Moment_Gyroscope.pdf, Accessed 23 March 2011
35. Defendini, A., Lagadec, K., Guay, P., Blais, T., Griseri, G., "Low cost CMG-based AOCS designs", Proc. 4th International Conf. on Spacecraft Guidance, Navigation and Control Systems, pages 393-398, 2000
36. A. Bradford, "BILSAT-1: A Low Cost, Agile, Earth Observation Microsatellite for Turkey" International Astronautical Federation, October, Houston, USA, 2002
37. ATV Data Sheet, <http://www.astrium.eads.net/en/programme/atv.html>, Accessed 23 March 2011
38. Wertz, J. R., Larson, Space Mission Analysis & Design, Microcosm, Torrance, California, 1999
39. SSTL Sun Sensor Data Sheet, www.sstl.co.uk/Downloads/Datasheets/Sun-sensor, Accessed 29 March 2011
40. Bradford Sun Sensor, www.bradford-space.com/pdf/be_datasheet_sun_sep2006.pdf, Accessed 29 March 2011
41. Berlin, P., Satellite Platform Design, Department of Space Science, University of Lulea, 5th Edition, 2007

José Meseguer, Isabel Pérez-Grande, Angel Sanz-Andrés
and Gustavo Alonso

One of the problems that needs to be solved in order to achieve a successful space mission is to ensure suitable thermal behavior of all the spacecraft subsystems, which may not seem critical or problematic in the case of Earth-based equipment. However, it is crucial in the space environment. The physical and technical basis for the thermal control design of spacecraft is the main subject of this chapter.

This chapter is mainly devoted to the thermal control subsystem, whose task is to maintain the temperature of all spacecraft components, subsystems, engineering equipment, payloads and the total flight system at safe operating and survival levels during the entire life of the spacecraft for all mission phases. Like any other subsystem, the spacecraft thermal control subsystem is essential to ensure reliable operation and long-term survival of any spacecraft.

The thermal control process generally involves the controlled exchange of heat between the different parts of the vehicle, and between the vehicle and the environment driven by thermal radiation. The latter can be seriously affected if the surface properties (solar absorptance α , and emissivity ϵ) are modified by environmental conditions. In general, the environmental effects related to vacuum, neutral environment, radiation and micrometeoroids/orbital-debris, modify the absorptance/emissivity ratio of the spacecraft's external surfaces, whereas those related to plasma affect the re-attraction of contamination.

In the next section, the fundamentals of heat transfer for spacecraft thermal design are presented. A brief review of the different technologies used for spacecraft thermal control is presented in Sect. 13.2. Finally, Sect. 13.3 is devoted to describing the main aspects of thermal control design, analysis, and testing.

J. Meseguer · I. Pérez-Grande · A. Sanz-Andrés · G. Alonso (✉)
Universidad Politécnica de Madrid Madrid, Spain
e-mail: gustavo.alonso@upm.es

13.1 Fundamentals of Heat Transfer for Spacecraft Thermal Design

In this section, some basic concepts about heat conduction and thermal radiation heat transfer are presented. Owing to the vacuum conditions in the space environment, convection is not considered. The aim of this short review is to make the reader familiar with the concepts and notation that will be used in this chapter. For more detailed information about heat transfer in general, a number of general textbooks on heat transfer, such as [1–3] are available. For more specific information about thermal radiation, the reader is referred to [4, 5].

13.1.1 Conductive Heat Transfer

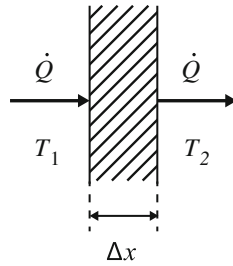
Heat conduction is the transfer of thermal energy between regions of matter in response to a temperature gradient. The physical mechanism is related to the transfer of free electrons from areas with higher energy (higher temperature) to areas with lower energy (lower temperature), and to lattice vibration. Thus, heat conduction requires the presence of molecules (matter); it is not possible in a vacuum, and takes place within solids and fluids (liquids and gases).

The equation that describes heat conduction is Fourier's law. It allows the calculation of heat fluxes for a given temperature field. Temperatures are calculated from the principle of energy conservation.

13.1.1.1 Fourier's Law

Fourier's law is an empirical law derived from experimental evidence and observation. Consider a wall of thickness Δx and area A , as shown in Fig. 13.1. Let the temperature be uniform over the area A on both wall surfaces. Assume that the temperature is higher on the left face of the wall and lower on the right one.

Fig. 13.1 One-dimensional heat conduction, \dot{Q} , across a *solid wall* of thickness, Δx , whose surfaces are at temperatures T_1 and T_2 , respectively



Fourier's law states that the rate of heat flux, \dot{Q} , through a uniform material is directly proportional to the area of heat transfer and to the temperature gradient, ΔT , in the direction of the heat flux, and is inversely proportional to the length of the path flow, Δx . Thus

$$\dot{Q} \propto A \frac{\Delta T}{\Delta x}. \quad (13.1)$$

The constant of proportionality is the so-called thermal conductivity. Hence

$$\dot{Q} = kA \frac{\Delta T}{\Delta x}. \quad (13.2)$$

The thermal conductivity, k , is a physical property that is a characteristic of the materials, and its SI units are $\text{W}/(\text{m} \cdot \text{K})$. It is a measure of how fast heat flows in the material.

The thermal conductivity of different materials varies considerably. Thus, there are up to four orders of magnitude of difference between the thermal conductivity of gases and that of conductive metals. Table 13.1 shows a list of the thermal conductivity of several materials used in spacecraft design.

Evaluating Eq. 13.2 in the limit $\Delta x \rightarrow 0$, the heat rate is given by

$$\dot{Q} = -kA \frac{dT}{dx}. \quad (13.3)$$

Note that the negative sign in the previous equation indicates that the transfer of heat occurs from higher to lower temperatures. The heat flux density, $q = (\dot{Q}/A)$, or the heat flux per unit of time and area is

$$q = -k \frac{dT}{dx}. \quad (13.4)$$

Fourier's law has been introduced under the restricted and simplified conditions of one-dimensional, steady-state conduction in a plane wall. In these conditions, the temperature distribution can be shown to be linear. However, Fourier's law also applies to multi-dimensional and transient conduction in complex geometries. In these cases the temperature field is not evident. Thus, a more general form of Fourier's law for a three-dimensional case can be written as

$$\mathbf{q} = -k \nabla T = -k \left(\frac{\partial T}{\partial x} \mathbf{i} + \frac{\partial T}{\partial y} \mathbf{j} + \frac{\partial T}{\partial z} \mathbf{k} \right). \quad (13.5)$$

Note that the heat flux, \mathbf{q} , is a vector quantity and $T(x, y, z)$ is the scalar temperature field. It is implicit in Eq. 13.5 that the heat flux vector, \mathbf{q} , is perpendicular to the isothermal surfaces.

In the one-dimensional case, mainly in the field of thermal modeling, it is sometimes useful to write Fourier's law in terms of thermal resistance, R_{th} , a magnitude that depends not only on the material but also on the geometry. It is defined as $R_{\text{th}} = d/(k \cdot A)$, in this case d being the heat path length, and A the area of heat flux. Substituting this definition into Eq. 13.2 yields $\dot{Q} = (\Delta T/R_{\text{th}})$. This means that a heat flux \dot{Q} can be analyzed in a similar way to an electric current. In Ohm's law, \dot{Q} would be the intensity, the temperature difference ΔT would correspond to the electrical voltage, and the thermal resistance, R_{th} , to the electrical resistance.

13.1.1.2 The Heat Diffusion Equation

Fourier's law described in the previous subsection allows the calculation of heat fluxes for a given temperature field. However, one of the major objectives in a conduction analysis is to determine the temperature field in a domain as a result of the conditions imposed on its boundaries. Heat fluxes can then be calculated from this temperature field.

To do this, the energy balance equation applied to an elemental volume can be stated, in Cartesian coordinates as

$$\frac{\partial}{\partial x} \left(k \frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left(k \frac{\partial T}{\partial y} \right) + \frac{\partial}{\partial z} \left(k \frac{\partial T}{\partial z} \right) + \dot{q}_v = \rho c_p \frac{\partial T}{\partial t} \quad (13.6)$$

where \dot{q}_v is the rate at which energy is generated per unit volume of the medium, ρ is the medium density, c_p is the thermal capacity, and $\rho c_p (\partial T / \partial t)$ is the time rate of change of the internal energy of the medium per unit volume. For the one-dimensional case, if the thermal conductivity is constant and $\dot{q}_v = 0$, then Eq. 13.6 can be simplified and written in the form

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{\alpha_d} \frac{\partial T}{\partial t} \quad (13.7)$$

where $\alpha_d = k/(\rho c_p)$ is the thermal diffusivity.

13.1.1.3 Boundary and Initial Conditions

To determine the temperature field in a medium it is necessary to solve the heat diffusion equation. To do that, it is necessary to know some physical conditions at the boundaries. These can be given as temperatures, heat fluxes, or a combination of both. If the problem is time dependent, then

Table 13.1 Thermal conductivity, k , of various materials at room temperature [6]

Group	Material	Chemical composition	k [W/(m · K)]
Aluminum and Al alloys	Aluminum (ISO Al 99.5)	99.5 % Al	230
	Aluminum–Copper alloy (ISO AlCu4Mg1)	4.5 % Cu, 1.5 % Mg, 0.6 % Mn, remaining Al	150–180
	Aluminum–Magnesium alloy (ISO AlMg2)	1.7–2.4 % Mg, remaining Al	155
	Aluminum–Magnesium–Silicon alloy (ISO AlMgSi)	0.4–0.9 % Mg, 0.3–0.7 % Si, remaining Al	197–201
	Aluminum–Zinc alloy 7075	5.6 % Zn, 2.5 % Mg, 1.6 % Cu, 0.3 % Cr, remaining Al	134
	2219 Aluminum–Copper–Manganese alloy (ISO AlCu6Mn)	Cu 5.8–6.8 %, Mn 0.2–0.4 %, remaining Al	116–170
Copper and Cu alloys	Copper (Oxygen-free high-conductivity; OFHC)	99.95 % Cu	394
	Beryllium–Copper (CDA 170)	1.8 % Be, 0.3 % Co + Ni, remaining Cu	84–150
	Brass (α - β) leaded	40 % Zn, 2 % Pb, remaining Cu	117
	Phosphor Bronze (CDA 510)	5 % Sn, 0.2 % P, remaining Cu	75
Titanium and Ti alloys	Timetal 35A (IMI 115)	Commercially pure Ti	16
	Ti 6Al 4 V (IMI 318)	6 % Al, 4 % V, remaining Ti	6
	Ti 4Al 4Mo–Si (IMI 550)	4 % Al, 4 % Mo, 2 % Sn, 0.5 % Si remaining Ti	8
Stainless steels	Stainless steel A286	25 % Ni, 15 % Cr, 2 % Ti, 1.5 % Mn, 1.3 % Mo, 0.3 % V, remaining Fe	23.7
	Stainless steel AISI 304L	8–12 % Ni, 18–20 % Cr, 2 % Mn max, 1 % Si max, 0.03 % C max, remaining Fe	16.2
	Stainless steel AISI 316L	12 % Ni, 17 % Cr, 2.5 % Mo, 2 % Mn, 1 % Si, 0.03 % C max, remaining Fe	16
Miscellaneous metallic materials	Magnesium–Aluminum–Zinc alloy	8.5 % Al, 0.5 % Zn, remaining Mg	90
	Magnesium–Aluminum–Zinc–Manganese alloy	3 % Al, 1 % Zn, 0.2 % Mn, remaining Mg	84
Adhesives, coatings and varnishes	Araldite AV138/HV998 (100/40 pbw)	Epoxy	0.35
	D.C. 93500	Silicone	0.146
	Eccobond ‘solder’ 56C	Epoxy–Silver-loaded	5.8
	RTV S 691	Silicone, filled	0.39
	RTV S 695	Silicone	0.21
	Epo-tek 930	2-part Epoxy, Boron nitride filled	4.1
Potting compounds, sealants and foams	D.C. 340	Silicone compound, filled	0.55
	RTV 566	Silicone (methyl, phenyl)	0.3
	Stycast 1090	Epoxy	0.167
	Stycast 2850FT	Epoxy	1.44
	Upilex foam	Polyimide	0.03

(continued)

Table 13.1 (continued)

Group	Material	Chemical composition	k [W/(m · K)]
Reinforced plastics	Makrolon GV 30	Polycarbonate/glass	0.16
Rubbers and elastomers	Eccoshield SV-R	Metal-filled Silicone	4.3
Thermoplastics (non-adhesive tapes and foils)	Sheldahl 146368	Fluorocarbon (FEP), Silver and Inconel coated	0.194
	Sheldahl 146372	Fluorocarbon (FEP), aluminized	0.194
	Sheldahl 146633	Polyimide Kapton HN, Aluminum and ITO coated	0.155
	Sheldahl G423020	Fluorocarbon (FEP), aluminized and ITO coated	0.194
	Dunmore DE291	Polyimide (Kapton HN), aluminized with protective coating on both sides	0.028
	Dunmore DM100	Polyimide (Kapton HN), aluminized/acrylic adhesive	0.0155
	Dunmore DE 028	Polyethylene Terephthalate/PETP, aluminized	0.61
	Dunmore DE 320	Polyimide (Kapton HN), aluminized	0.155
	Dunmore TM05564	Fluorocarbon (FEP-Type C), aluminized	0.194
	Kapton H, HN	Polyimide	0.155
	Hostaform C9021	Acetal copolymer	0.31
	PETP (Mylar, Melinex, Terphane,...)	Polyethylene Terephthalate	0.61
	PTFE (Teflon, Halon, Fluon, Hostafflon)	Polytetrafluoroethylene	0.25
	Sheldahl 146401 (previously G401500)	Fluorocarbon (FEP), Silver and Inconel Coated	0.194
	Sheldahl 146383 (previously G400900)	Fluorocarbon (FEP), aluminized	0.194
	Sheldahl 146631 (previously G425120)	Polyimide (Kapton H), ITO/aluminized	0.155
UPILEX S	Polyimide	0.29	
Thermoset plastics	Rexolite 1422	Polystyrene, cross-linked	0.146

the conditions existing in the medium at some initial time must also be provided. Mathematically, the heat diffusion equation is a differential equation that requires integration constants in order to have a unique solution. Boundary conditions are in fact the mathematical expressions or numerical values necessary for this integration.

13.1.1.4 Conductive Shape Factors

In two- or three-dimensional conduction problems where only two temperature levels are involved, a conductive shape factor, S_c , can be defined in such a way that the heat transfer rate may be expressed as

$$\dot{Q} = S_c k \Delta T. \quad (13.8)$$

The conductive shape factor, S_c , has been obtained analytically for numerous two and three dimensional systems and the values for some simple configurations can be found in [3, 7].

13.1.1.5 Numerical Methods in Heat Conduction

The equations shown in the previous sections can be solved by analytical methods only in certain particular cases. These solutions are available in the literature for different geometries. When it is not possible to obtain an exact mathematical solution, the best alternative is often to use a numerical technique. Numerical solutions allow temperature determination only at discrete points of a system, unlike analytical solutions, which give the temperature field for all points of the medium. The most widely used numerical methods for heat conduction are the finite-difference, finite-element, and boundary-element methods. For thermal control purposes lumped capacitance models are used. To apply numerical methods, the system has to be divided into regions where the temperature is assumed to be uniform, this temperature value being the average temperature of the region. Generally, in order to assign properties, the centers of these regions are used as reference points. They are called nodes. The meshing of the system, that is,

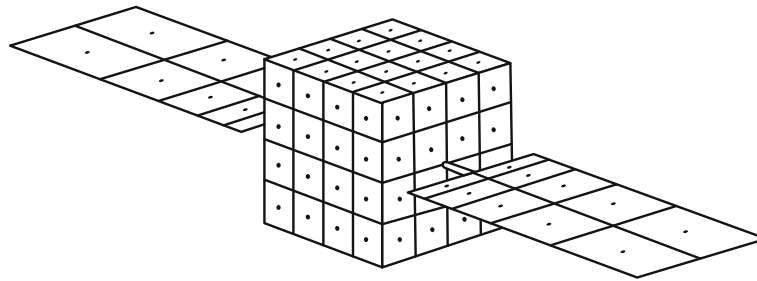


Fig. 13.2 Example of a typical discretization of a satellite surface for a preliminary thermal control analysis, with the nodes indicated by *black dots*

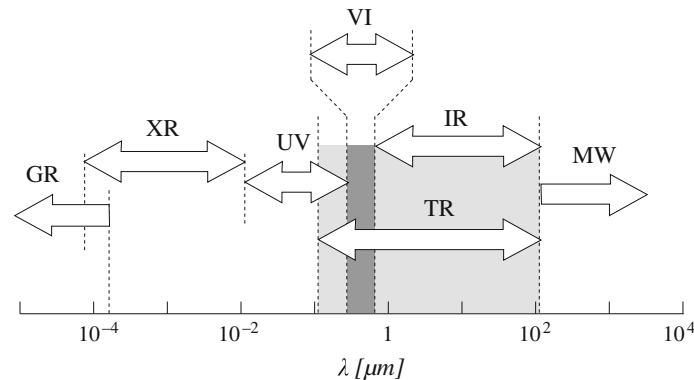


Fig. 13.3 Electromagnetic spectrum classification according to radiation wavelength, λ , showing the wavelength ranges corresponding to thermal radiation; *GR* gamma rays; *XR* X-rays; *UV* ultraviolet; *VI* visible; *IR* infrared; *TR* thermal radiation; *MW* microwaves

the selection of the nodes, is done arbitrarily based on experience. The accuracy of the calculations strongly depend on the number of nodes, their sizes, and their locations. The higher the number of nodes, the more accurate the results, but the procedure is more time consuming. Figure 13.2 shows a typical discretization of a satellite surface.

13.1.2 Thermal Radiation Heat Transfer

Thermal radiation is electromagnetic radiation emitted from all matter that is at a non-zero absolute temperature, in the wavelength range from 0.1 to 100 μm . It includes part of the ultraviolet (UV) range, and all the visible and infrared (IR) ranges. It is called thermal radiation because it is caused by and affects the thermal state of matter. Figure 13.3 shows the electromagnetic spectrum with the region of thermal radiation indicated on it. Note that this is a simplified version of Fig. 3.3.

Therefore, thermal radiation does not require a material medium for its propagation. Although in the context of spacecraft thermal design the interest in radiation is mainly focused on solid surfaces, emission may also occur from

liquids and gases. The mechanism of radiation emission is related to the energy released as a result of oscillations or transitions of the electrons that constitute matter. These oscillations are sustained by the internal energy, and therefore the temperature of the matter.

All forms of matter emit radiation since they are at non-zero absolute temperature. For gases and for semi-transparent media, thermal radiation is a volumetric phenomenon. This can be of interest when studying the behavior of lenses, for instance, incorporated into optical devices.

Since thermal radiation is electromagnetic radiation, the properties of the propagation of electromagnetic waves can be applied. The most relevant ones are the frequency, ν , and the wavelength, λ , which are related through $\lambda = (c/\nu)$, where c is the speed of light in the medium.

The spectral nature of thermal radiation is one of the two features that make its study quite complex. The second feature is related to its directionality. A surface may have certain directions with preferential emission; in this case, the distribution of the emitted radiation is directional. When the radiative properties do not depend on the direction, the surface is said to be ‘diffuse’.

The thermal radiation emitted by a surface will strike other surfaces and will be partially reflected, partially

Fig. 13.4 Thermal radiation interactions on a surface: Φ_e emitted radiation; Φ_i incident radiation; Φ_a absorbed radiation; Φ_r reflected radiation; Φ_t transmitted radiation

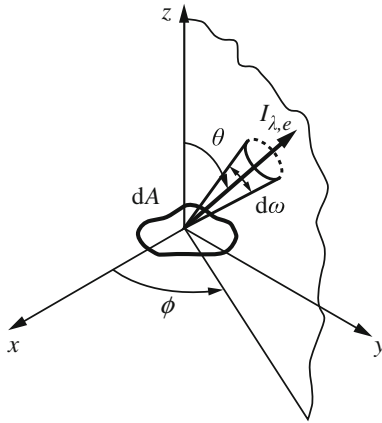
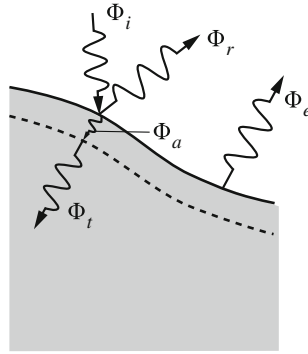


Fig. 13.5 Intensity of radiation, I_e , emitted in the direction (θ, ϕ) by an elemental surface dA (contained in the x - y plane), with $d\omega$ being the solid angle unit about this direction

absorbed, and partially transmitted. Figure 13.4 shows the different thermal radiation interactions on the surface of a body. The symbol Φ in the figure stands for the radiant energy per unit time. As can be seen, the surface emits Φ_e , receives the incident radiation Φ_i , part of which, Φ_a is absorbed, Φ_r is reflected, and Φ_t is transmitted.

The intensity of emitted radiation, I_e , is defined as the rate at which the radiant energy, $\delta\Phi$, is emitted at the wavelength λ in the direction (θ, ϕ) , per unit area of the emitting surface normal to this direction, per unit solid angle $d\omega$ about this direction, and per unit wavelength interval $d\lambda$ about λ , as indicated in Fig. 13.5. Thus the spectral intensity is

$$I_{\lambda,e}(\lambda, \theta, \phi, T) = \frac{\delta\Phi_e}{dA \cos\theta d\omega d\lambda} \quad (13.9)$$

In order to obtain the thermal interactions in all directions and wavelengths, the intensity of the radiation is successively integrated. Thus, the spectral hemispherical emissive power E_λ is the rate at which radiation of wavelength λ is emitted in all directions from a surface per unit wavelength interval $d\lambda$ about λ and per unit surface area. It has the form

$$E_\lambda(\lambda, T) = \int_0^{2\pi} \int_0^{\pi/2} I_{\lambda,e}(\lambda, \theta, \phi, T) \cos\theta \sin\theta d\theta d\phi \quad (13.10)$$

where the solid angle, $d\omega$, has been written as $d\omega = \sin\theta d\theta d\phi$, according to the spherical coordinates defined in Fig. 13.5.

Finally, by integrating Eq. 13.10 in all wavelengths, the total emissive power, E , is obtained as

$$\begin{aligned} E(T) &= \int_0^\infty E_\lambda(\lambda, T) d\lambda \\ &= \int_0^\infty \int_0^{2\pi} \int_0^{\pi/2} I_{\lambda,e}(\lambda, \theta, \phi, T) \cos\theta \sin\theta d\theta d\phi d\lambda. \end{aligned} \quad (13.11)$$

The previous definitions, Eqs. 13.9–13.11, refer to the radiation emitted by a surface. Analogous definitions and mathematical expressions can be established for the incident radiation on a surface, called irradiation, G , and for all the radiation leaving a surface (the sum of the reflected radiation and the emitted radiation), called radiosity, J . Both can be defined at a spectral and directional level, at a spectral hemispherical level, and as a total magnitude integrated over all directions and all wavelengths.

13.1.2.1 Black-Body Radiation

A black-body is characterized by an ideal surface that absorbs all incident radiation, at all wavelengths and all directions. Therefore, it is the perfect absorber. As a consequence of this definition, the black-body has three properties: (a) it is the surface that most emits for a given temperature and wavelength, (b) black-body radiation does not depend on the direction, that is, black-body radiation is diffuse, and (c) total black-body radiation in a vacuum only depends on its absolute temperature.

Since the black-body is the perfect absorber and emitter, it will be used as a reference to compare the radiative properties of real surfaces.

The spectral emissive power $E_{\lambda,b}$ of a black-body was obtained by Planck as

$$E_{\lambda,b}(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 (e^{hc/(\lambda kT)} - 1)} \quad (13.12)$$

where $h = 6.626 \times 10^{-34} \text{ J} \cdot \text{s}$ is the Planck constant, $k = 1.380 \times 10^{-23} \text{ J/K}$ is the Boltzmann constant, and $c = 2.998 \times 10^8 \text{ m/s}$ is the speed of light in vacuum. The subscript b stands for black-body. This expression is the Planck distribution. The graphical representation of constant temperature curves, Fig. 13.6, provides valuable information.

From Fig. 13.6, it can be seen that for a given wavelength the emitted radiation increases with temperature. Each constant temperature curve has a maximum. This

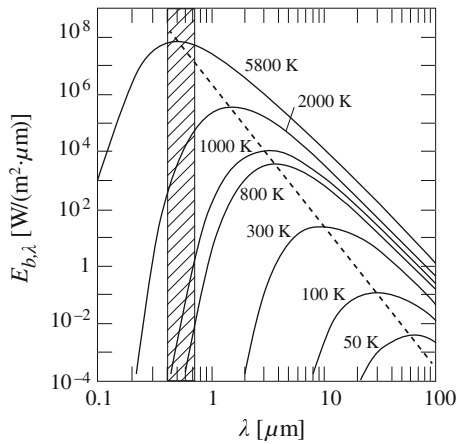


Fig. 13.6 Spectral emissive power of a black-body, $E_{\lambda,b}$, versus wavelength, λ , according to Eq. 13.12. Figures in the curves indicate the black-body temperature, T . The striped area indicates the visible spectral region

maximum moves towards longer wavelengths as the temperature decreases. If Eq. 13.12 is derived to obtain the wavelength where the maximum emissive power occurs, the result is Wien's displacement law, given by $\lambda_{max}T = C_3$, where $C_3 = 2,898 \mu\text{m} \cdot \text{K}$. Furthermore, the region where radiation is concentrated also moves towards longer wavelengths as the temperature decreases. As discussed in Sect. 3.2.3, the Sun's radiation can be plotted as equivalent to a black-body at approximately 5,781 K, and it is evident that the visible part of the spectrum is included in it. For temperatures of about 300 K, similar to the Earth's surface, radiation is concentrated in the infrared part of the spectrum.

The spectral emissive power of a black-body is given in Eq. 13.12, which can be integrated to obtain the total emissive power of a black-body giving

$$E_b(T) = \int_0^{\infty} E_{\lambda,b}(\lambda, T) d\lambda = \sigma T^4 \quad (13.13)$$

where $\sigma = 5.67 \times 10^{-8} \text{ W/(m}^2 \cdot \text{K}^4)$ is the Stefan-Boltzmann constant. This result, $E_b = \sigma T^4$, is known as the Stefan-Boltzmann law, and can be used to obtain the radiation emitted by a black-body in all directions and all wavelengths.

When it is necessary to know the fraction of energy emitted by a black-body at a temperature T within the bandwidth between λ_1 and λ_2 , it can be calculated from

$$F_{\lambda_1 \rightarrow \lambda_2} = \frac{1}{\sigma T^4} \int_{\lambda_1}^{\lambda_2} E_{\lambda,b}(\lambda, T) d\lambda. \quad (13.14)$$

The quantity $F_{0 \rightarrow \lambda}$, that is, the fraction of energy emitted by a black-body between 0 and the wavelength λ depends only on the product λT . It can be found tabulated in the references of heat transfer and thermal radiation mentioned previously.

13.1.2.2 Properties of Real Surfaces

The black-body has been defined as an ideal surface to be used as a reference to describe the behavior of real surfaces. Since the black-body is the perfect emitter, any real surface will emit less than the black-body at the same temperature, same wavelength and same direction. Thus, the spectral directional emissivity is defined as the ratio between the real emission and the black-body emission for the same temperature, wavelength and direction

$$\varepsilon(\lambda, \theta, \phi, T) = \frac{I_{\lambda,e}(\lambda, \theta, \phi, T)}{I_{\lambda,b}(\lambda, T)}. \quad (13.15)$$

Spectral hemispherical emissivity is defined as the ratio

$$\varepsilon(\lambda, T) = \frac{E_{\lambda}(\lambda, T)}{E_{\lambda,b}(\lambda, T)}. \quad (13.16)$$

When the surface is diffuse, the spectral hemispherical emissivity has the same value as the spectral directional emissivity. When the properties of the surface depend on the direction, the spectral hemispherical emissivity can be obtained by appropriately integrating Eq. 13.15 according to the definition given in Eq. 13.16 [2].

The second group of radiant properties is related to irradiation. As already said, a surface will be irradiated by the radiation coming from other surfaces. This incident radiation will be partially reflected, partially absorbed and partially transmitted (see Fig. 13.4). Based on this fact, three radiative properties are defined. First, the spectral directional absorptance is defined as the fraction of the incident radiation that is absorbed for a given direction and wavelength. Thus

$$\alpha(\lambda, \theta, \phi, T) = \frac{I_{\lambda,i,abs}(\lambda, \theta, \phi, T)}{I_{\lambda,i}(\lambda, \theta, \phi)}. \quad (13.17)$$

Second, the spectral directional reflectance is defined as the fraction of the incident radiation that is reflected for a given direction and wavelength. In this case

$$\rho(\lambda, \theta, \phi, T) = \frac{I_{\lambda,i,reflec}(\lambda, \theta, \phi, T)}{I_{\lambda,i}(\lambda, \theta, \phi)}. \quad (13.18)$$

Third, the spectral directional transmittance is defined as the fraction of the incident radiation that is transmitted for a given direction and wavelength. Thus

$$\tau(\lambda, \theta, \phi, T) = \frac{I_{\lambda,i,trans}(\lambda, \theta, \phi, T)}{I_{\lambda,i}(\lambda, \theta, \phi)}. \quad (13.19)$$

In the same way as was done for the emissivity, these coefficients can be integrated to obtain their values for all directions and for all wavelengths. Therefore, the spectral hemispherical absorptance is defined as

$$\alpha(\lambda, T) = \frac{G_{\lambda,abs}(\lambda, T)}{G_{\lambda}(\lambda)} \quad (13.20)$$

the spectral hemispherical reflectance is defined as

$$\rho(\lambda, T) = \frac{G_{\lambda,reflec}(\lambda, T)}{G_{\lambda}(\lambda)} \quad (13.21)$$

and the spectral hemispherical transmittance is defined as

$$\tau(\lambda, T) = \frac{G_{\lambda,trans}(\lambda, T)}{G_{\lambda}(\lambda)}. \quad (13.22)$$

For each of the levels of definition of these coefficients (spectral directional and spectral hemispherical) the following relationship $\alpha + \rho + \tau = 1$ is verified. For opaque surfaces the transmittance is zero, and hence $\alpha + \rho = 1$.

Kirchhoff's law establishes that $\alpha(\lambda, \theta, \phi, T) = \varepsilon(\lambda, \theta, \phi, T)$; that is, for each direction and wavelength the emissivity equals the absorptance. If the surface is diffuse, then $\alpha(\lambda, T) = \varepsilon(\lambda, T)$ can be derived from Kirchhoff's law.

A surface is defined as 'gray' when its properties are independent of the wavelength, or more particularly that $\alpha(\lambda, T) = \alpha(T)$ and $\varepsilon(\lambda, T) = \varepsilon(T)$. Most real surfaces are not exactly gray, but the equations will still be valid if the properties do not change with the wavelength in the range of interest, that is, in the range of wavelengths where radiation exchange takes place.

13.1.2.3 View Factors

The view factor, F_{ij} , between two surfaces, A_i and A_j , is defined as the fraction of the radiation leaving surface i that reaches surface j . It is also termed the configuration factor or the geometrical factor. Mathematically, the view factor between two diffuse infinitesimal surfaces, dA_i and dA_j , with uniform radiosity can be expressed as

$$dF_{ij} = \frac{\cos \theta_i \cos \theta_j}{\pi r^2} dA_j \quad (13.23)$$

where r is the distance between both elements, and θ_i and θ_j are the angles between the normal vector to each surface and the line of sight between the elements (Fig. 13.7).

From the mathematical definition of the view factor, the reciprocal relation, $A_i F_{ij} = A_j F_{ji}$, can be obtained. This expression is very useful in determining one view factor when the reciprocal is known.

When a set of n surfaces forms an enclosure, the summation rule of view factors applies to any of the surfaces. It is given by

$$\sum_{j=1}^n F_{ij} = 1. \quad (13.24)$$

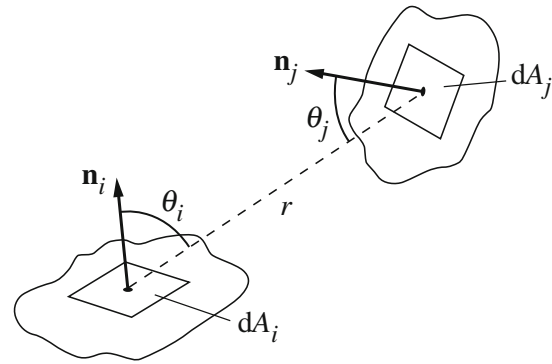


Fig. 13.7 Magnitudes to calculate the view factor between two elemental surfaces, dA_i and dA_j , placed a distance r apart. \mathbf{n}_i , \mathbf{n}_j , are unity vectors normal to the surfaces

Note that the term F_{ii} , the view factor of a surface with respect to itself, may be non-zero if the surface is concave, that is, if the surface can see itself.

The view factors for very simple geometries can be analytically calculated, although more complex geometries would require the use of numerical methods (in fact, the calculation of view factors is a standard feature of most commercial tools dealing with spacecraft thermal control). The view factors for a number of geometrical configurations have been already calculated and are available in the literature [5, 7].

13.1.3 In-Orbit Thermal Loads

Spacecraft in orbit receive thermal radiation from three primary sources, reflect part of it back and radiate energy to the cold sink of space. The determination of these external thermal loads absorbed by a given spacecraft is an essential preliminary task to be carried out when the energy equation needs to be solved to predict the system temperatures. Thus, the change of energy of the spacecraft has to be equal to the sum of the power dissipated by the electronics plus the absorbed external thermal loads, minus the energy emitted by the spacecraft to the outer space cold sink. Hence, as introduced in Chap. 3, and as shown in Fig. 13.8, the three main sources of incoming energy in a spacecraft are the solar radiation, the albedo radiation, and the planetary infrared radiation. In some missions, other sources of external heating also have to be taken into account, as is the case of the aerodynamic heating that occurs when a space vehicle passes through an atmosphere. This happens, for instance, during the aerobraking phase of a mission to a planet that possesses an atmosphere. In such cases, these aerodynamic thermal loads have to be carefully analyzed. Furthermore, since the design of the thermal control subsystem has to meet the requirements of all mission phases,

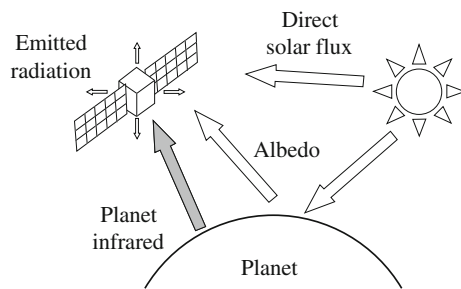


Fig. 13.8 Thermal interactions of a space vehicle with the environment: solar radiation, albedo radiation, planet infrared radiation, and spacecraft emitted radiation

from the launch until the end of the operating lifetime, the atmospheric portion of the launch phase, though short, also needs to be analyzed. In the following, only the three main sources of external heating are described. The reader is referred to [8] for a detailed description of atmospheric thermal loads.

13.1.3.1 Solar Radiation

Direct solar radiation is the main source of heating and power of most spacecraft. The term solar constant, or solar irradiance, is defined as the electromagnetic radiation from the Sun that falls on a unit area of surface normal to the line from the Sun, per unit time, outside the atmosphere, at one astronomical unit (au) [9]. According to [10], and as introduced in Chap. 3, its value is $G_s = 1,366.1 \text{ W/m}^2$. A typical margin of $\pm 10 \text{ W/m}^2$ [11] is applied for thermal calculations to take solar fluctuations into account. These fluctuations are due to the natural variability of the solar output during the Sun's eleven-year cycle, to the slightly elliptical orbit of Earth, and to measurement uncertainties. This value of the solar constant is the integrated value over the spectrum. For thermal purposes, the spectral distribution of the solar radiation can be considered as a black-body at 5,781 K. The real irradiation spectral distribution at 1 au can be found in Fig. 3.3, taken from [12]. This spectral distribution fits quite well with the emissive power of a black-body at approximately 5,781 K at the Earth's distance from the Sun.

Looking in detail at this distribution, it can be seen that 99 % of solar radiation is between 0.15 and 10 μm . The visible part of the spectrum is included within this range, and it represents 46 % of the total radiation. A further 47 % is in the near-infrared range and 7 % is ultraviolet radiation. The maximum is at 0.45 μm .

At 1 au the apparent diameter of the Sun is 0.5° . This means that at Earth orbits, for thermal calculation purposes, solar radiation can be considered to be parallel rays. In the case of thermal analysis performed for closer distances to the Sun, as in the case of a thermal study of a Mercury

orbiter, the effect of the solar angle may have to be taken into account; mainly for optical devices.

When the solar constant, G_s , has to be determined at a distance to the Sun, $d = d_{sc-S}$, different from 1 au, it can be derived from the solar constant defined at 1 au and the distance from the Earth to the Sun d_{E-S} . Applying the conservation of solar power passing through concentric spherical surfaces of different diameters, d_{sc-S} , gives

$$G_s(d_{sc-S}) = G_s(1 \text{ au}) \left(\frac{d_{E-S}}{d_{sc-S}} \right)^2. \quad (13.25)$$

Values of the solar constant at different planetary orbits can be found in Table 4.3.

The calculation with a simple analytical expression of the solar radiation absorbed by a flat surface of area A , whose normal vector forms an angle θ with the solar rays is

$$\dot{Q}_{Sun} = \alpha G_s A \cos \theta \quad (13.26)$$

where α is the solar absorptance of the surface.

13.1.3.2 Albedo Radiation

Albedo is the fraction of incident solar radiation which is reflected off a planet. Therefore its influence as a thermal load is higher for low altitude orbits, and particularly for low Earth orbits (LEO). The albedo coefficient, α , may be highly variable over the planet's surface, as happens on Earth's surface. Oceans absorb most of the incident radiation, the local albedo being between 0.05 and 0.10. Ice or snow, for example the Antarctic surface, reflects most of the solar radiation and the local albedo coefficient is about 0.95. In continental areas the albedo can range from small values over forests to higher values over desert areas. The presence of clouds, mainly the quantity and the type, is also an important factor that alters the local albedo. An albedo value of 0.8 is typical for cloudy areas. For thermal design purposes of low orbit satellites, mean values can be used because the changes occur rapidly. The mean value for Earth is taken as approximately 0.3 [9]. Geometric and Bond albedo values (defined in Chap. 4) for each of the planets and for the Moon can be found in Table 4.4.

Due to the roughness of a planet's surface, the albedo is assumed to be diffuse. As an approximation, the spectral distribution of the reflected light is considered to be the same as that of incident light.

When determining the thermal loads on a spacecraft, albedo loads are only applicable when the portion of the planet that is seen by the spacecraft is sunlit. The calculation is often a complex task usually carried out with the help of computer tools. As it applies only to the portion of the planet illuminated by the Sun, its value will depend on the solar zenith angle, that is, the angle between the Sun-planet

vector and the planet-spacecraft vector. The orbit angle, β , the minimum angle between the spacecraft's orbit plane and the Sun-Earth vector, also has to be taken into account. For Earth orbits, detailed information on this data and the corrections that have to be applied to those angles to calculate the albedo coefficient is presented in [13], based on measurements of the NOAA and ERBS satellites.

For simplified analytical estimations, the albedo absorbed energy on a surface of area A can be calculated assuming that the planet behaves as a reflecting sphere

$$\dot{Q}_{alb} = aG_s A F_{sc-p} \cos \phi \quad (13.27)$$

for $-\pi/2 \leq \phi \leq \pi/2$, where a is the albedo coefficient, G_s is the solar constant, ϕ is the solar zenith angle, and F_{sc-p} is the view factor between the surface and the planet. The angle ϕ takes into account the fact that the albedo is at a maximum at the sub-solar point and it becomes zero when the planet seen by the spacecraft is in eclipse. Other simple analytical models that take into account seasonal effects and latitude and longitude on the Earth's surface can be found in [9].

In the case of Earth orbits, the albedo loads are relevant only for low altitudes. For telecommunications satellites in geostationary orbits (GEO) these loads are practically negligible.

13.1.3.3 Planetary Radiation

Planetary radiation is the thermal radiation emitted by a planet. It is also called outgoing long-wave radiation. It is a combination of the radiation emitted by the planet's surface and by the atmospheric gases. It is diffuse radiation within the infrared part of the spectrum. As is the case with the albedo coefficient, the emission of a planet's surface varies from one point to another. For example, on Earth, it depends on the local time, on the presence of water (oceans), highly populated areas, desert areas, etc. Detailed information of these variations on the Earth's surface and correlations between the Earth's infrared radiation and the albedo coefficient are presented in [13].

Nevertheless, although planets are not strictly in thermal balance, the solar energy absorbed by a planet is almost balanced by the emitted radiation, a fact that can be used to determine a planet's radiative properties from the energy balance equation. For thermal purposes, the thermal energy emitted by planets can be characterized by means of the planet's black-body equivalent temperature. If the albedo coefficient, a , is known, a first estimation of the planet's equivalent black-body temperature can be obtained by equating the solar energy absorbed by the planet to the emitted energy

$$G_s \pi R_p^2 (1 - a) = 4\pi R_p^2 \sigma T_p^4 \quad (13.28)$$

where σ is the Stefan-Boltzmann constant, and R_p is the radius of the planet. Note that the resulting temperature is independent of the value of R_p , as can be seen in the

equation. Interestingly, for planets like Mercury with a long day compared to its year, the assumption of considering a single temperature to model the thermal behavior of the planet may not work properly, and more complex models are necessary. In the case of the Earth, with a mean albedo coefficient $a = 0.3$, its black-body equivalent temperature is $T_E = 254$ K. This corresponds with a flux of 240 W/m^2 on the Earth's surface. The black-body equivalent temperatures for other solar system planets are given in Table 13.2.

From the black-body temperature of the planet T_p , the planetary infrared thermal load on a spacecraft surface of area A can be calculated from

$$\dot{Q}_{planet} = \varepsilon A F_{sc-p} \sigma T_p^4 \quad (13.29)$$

where ε is the infrared emissivity of the spacecraft surface and F_{sc-p} is the view factor between the spacecraft surface and the planet.

As is the case with albedo loads, infrared planetary radiation is relevant primarily for low altitude orbits. Its influence on satellites in geostationary orbits is negligible.

13.2 Thermal Control Technologies

In the previous section, the theoretical laws and concepts that allow the formulation of the thermal balance of a spacecraft, considering it as either a whole system or divided into different subsystems, was presented. This thermal balance is one of the tasks to be recurrently performed by the thermal control team responsible for the thermal control design of a spacecraft.

Generally, the design process of the thermal control system will require the correction of unbalanced thermal loads, which is achieved by using suitable devices that allow for controlling and transmitting, or reducing heat fluxes between the different parts of the vehicle, and between the spacecraft and outer space. Such devices can be grouped under the general label of thermal control technologies, some of them being briefly described in this section (for more detailed information and data on these thermal control technologies, the reader is referred to [7, 14, 15]).

13.2.1 Thermal Control Surfaces

A satellite beyond the Earth's atmosphere is in an extreme situation with regard to temperature control. Conduction and convection are absent, and therefore radiative exchange alone determines the heat fluxes to and from the vehicle. Intense solar irradiation, radiative cooling to outer space, and internal heat generation determine the equilibrium temperature of a spacecraft. The balance between the solar

Table 13.2 The black-body equivalent temperatures at each of the planets and the Moon

	d (au)	R_p (km)	T_p (K)
Mercury	0.387	2,330	442
Venus	0.723	6,100	231.7
Earth	1.0	6,367.5	254
Mars	1.521	3,415	210.1
Jupiter	5.173	71,375	110.0
Saturn	9.536	60,500	81.1
Uranus	19.269	24,850	58.2
Neptune	30.034	25,000	46.6
Pluto	39.076	2,930	–
Moon	1.0	1,738	273

Distance to the Sun in au, d ; Radius of the planet, R_p ; Equivalent black-body temperature of the planet, T_p . Data concerning d from [7], and data concerning T_p from [9]

absorption and thermal emissivity of the surface is, therefore, crucial, in particular for autonomous parts directly exposed to solar radiation and thermally insulated from the main thermal mass of the spacecraft, such as instrument booms.

The thermal radiation from the satellite is a characteristic of its temperature, which is ideally close to room temperature, whose black-body radiation intensity maximum is at wavelengths close to 10 μm (Fig. 13.6).

Concerning spacecraft thermal design, the relevant characteristic of thermal radiation, illustrated in Figs. 13.6 and 3.3, is that solar radiation energy is concentrated in the short wavelength range, whereas room temperature emission is concentrated in the long wavelengths. Therefore, the wavelength gap between the maxima of the incoming and outgoing radiation offers the possibility of using optical selectivity for temperature control.

In space engineering, the radiation properties of a surface are characterized by its total normal solar absorptance, α (averaging across the ultraviolet spectrum to the near-infrared), and total hemispherical infrared emissivity, ε (averaging across the thermal infrared spectrum).

In a general sense, a coating consists of a layer (or layers) of any substance upon a substrate. Optical coatings have been used to control the temperature of spacecraft since the first successful orbital flight in 1957. Since then, coating materials have been developed to the point where reasonably stable coatings are available that give any desired value of the hemispherical total emissivity, ε , between 0.05 and 0.95 for any desired value of the solar absorptance, α , between 0.05 and 0.95.

When thermo-optical properties are considered (absorptance and emissivity), thermal control surfaces are classified into four basic types: solar reflector, solar absorber, total or flat reflector, and total or flat absorber.

Flat absorbers absorb throughout the spectral range, with relatively high solar absorptance and high emissivity. Among them are the so-called black paints. Paints which are flat absorbers can be made from black pigments such as the oxides or mixed oxides Cr_3O_4 , $\text{Fe}_2\text{O}_3\cdot\text{NiO}$, Fe_3O_4 , or $\text{Mn}_2\text{O}_3\cdot\text{NiO}$, grounded and dispersed in silicone elastomers or alkali metal silicate vehicles and applied to the base structure. Through anodizing, it is possible to achieve protection against corrosion as well as the desired optical properties.

Flat reflectors differ from flat absorbers in that they reflect energy throughout the spectral range (in both the solar and infrared regions). Flat reflectors are presently obtained with highly polished metals or with paints pigmented with metal flakes.

Solar reflectors are characterized by small values of the α/ε ratio. Solar reflectors reflect most incident solar energy while absorbing and emitting infrared energy. Such surfaces are useful for coating where low temperatures are needed. Typical materials belonging to this type are white paints and second surface mirrors.

Solar absorbers, absorb solar energy while emitting only a small percentage of infrared energy (that means high values of the α/ε ratio). Such materials absorb moderate amounts of solar energy striking their surfaces, but emit very small amounts of infrared radiation. Materials having these optical properties are not common.

The range of properties available for different types of materials is summarized in Fig. 13.9. The problem of selecting the specific coating for a given α/ε ratio is somewhat circumvented by the combination of two or more coatings in a checkerboard or stripe pattern to obtain the desired combination of average absorptance and emissivity.

Radiation properties of a large number of suitable materials for spacecraft thermal control can be found in [17, 18], and an extended list of coatings is presented in [14]. A summary of absorptance and emissivity values of some representative materials is given in Table 13.3.

13.2.2 Multilayer Insulations

A multilayer insulation (MLI), also called a thermal blanket, consists of several layers of closely spaced highly reflecting shields, which are placed perpendicular to the heat flow direction. The basic aim of MLI is to provide radiative insulation working as a multilayer radiative shield. Each internal layer is a very thin element (from 7 μm thickness) of a plastic material, typically Kapton® or Mylar®, coated with vacuum-deposited aluminum (VDA) or gold (VDG) on both sides for very low emissivity. This mirror-like aluminum finish is what makes the sheets highly

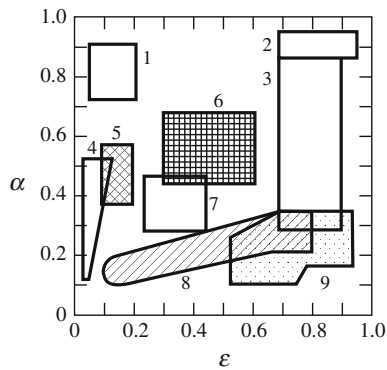


Fig. 13.9 Ranges of solar absorptance, α , and hemispherical total emissivity, ϵ , covered by different thermal control coatings: 1 selective black; 2 black paints; 3 gray and pastel paints; 4 polished metals; 5 bulk metals (*unpolished*); 6 sandblasted metals and conversion coatings; 7 metallic coatings; 8 dielectric films on polished metals; 9 white paints, second surface mirrors, metallized polymers [16]

reflecting and low emissive, which leads to a high resistance to radiative heat transfer between layers. The outer cover is usually thicker (from 125 μm) than the internal ones for increased mechanical strength. This external layer is usually aluminized only on its internal face because aluminum degrades when it is exposed to ultraviolet radiation and the α/ϵ ratio is too high. This configuration is called a ‘second surface mirror’. The external layer can be just bare Kapton or can be painted, for instance with a black carbon paint, in order to avoid undesired reflections. If better mechanical properties are needed, for example, to protect it from micrometeoroids, more resistant materials, such as Beta-cloth, can be used for this outer layer. In any case, since the outer cover may be exposed to solar radiation, its compatibility with ultraviolet radiation has to be carefully verified.

To avoid direct contact between shields, and therefore heat conduction between sheets, low-conductivity non-metallic spacers are used. In order to make the contact minimal, they are usually in the form of a mesh. Typical materials for this netting spacer are Dacron® (a brand name of polyethylene terephthalate, PET or PETE) and Nomex®. A typical cross section of an MLI is shown in Fig. 13.10.

The pile of layers is stacked together by adhesive closing flaps or stitches sewn with special non-metallic thread that has to be free of volatile components. Small Kapton pieces can be used to prevent the blanket from tearing due to tension in the threads. To close the lateral gaps and avoid the degradation of the internal layers, the outer cover is folded back on the internal layer. The blankets are attached to the spacecraft structure using hook-and-pile (i.e. Velcro®) fasteners and stand-off pins.

Proper venting of the MLI should be provided in order to avoid undue pressure loads on the shields during the ascent flight. Otherwise, the blanket would billow out like a balloon and the dynamic pressure could detach it from the

Table 13.3 Absorptance, α , and emissivity, ϵ , values of some representative materials

	α	ϵ
Carbon black paint NS-7 (black coating)	0.96	0.88
Catalac white paint (white coating)	0.24	0.90
Electrodag (conductive paint)	0.90	0.68
Black (anodized aluminum coating)	0.65	0.82
Aluminum polished (metal)	0.14	0.03
Aluminum (vapor deposited coating)	0.08	0.02

spacecraft. Thus, to allow proper venting of the blanket the sheets are perforated and/or sections of the edging are unsealed [19].

In order to prevent an accumulation of electrostatic charge and the resultant discharges, all the layers of the blanket have to be grounded to the spacecraft structure.

Detailed information about multilayer insulation materials, assembly and performance can be found in [7, 14, 20].

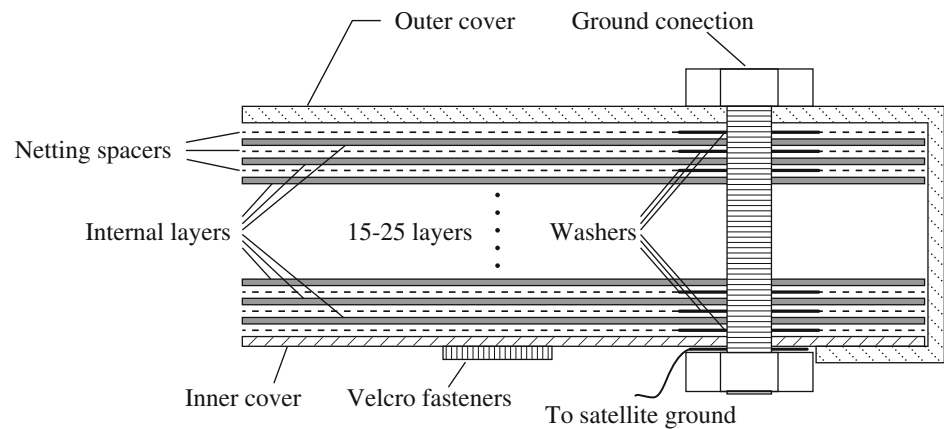
13.2.3 Radiators

Spacecraft heat is ultimately rejected to space by radiators, which are systems that take the waste thermal energy from a heat source and reject it by radiation to outer space through radiating surfaces. Although spacecraft radiators may adopt very different forms (from simple flat-plate radiators mounted on the side of a spacecraft to radiating panels deployed after the spacecraft is in orbit), in all cases the radiators reject heat by infrared radiation from their surfaces. Thus, the radiating power depends on the emissivity and temperature of the radiating surfaces. Obviously, the radiator must reject the spacecraft waste heat, but also any radiant-heat loads from the environment or other spacecraft surfaces that are absorbed by the radiator. Most radiators are therefore given surface finishes with high infrared radiation emissivity ($\epsilon > 0.8$) to maximize heat rejection, and low solar absorptance ($\alpha < 0.2$) to limit heat loads from the surroundings.

Depending on how the heat is transferred from the source to the radiating surfaces, radiators can be classified as passive radiators and active radiators. In the former, the connection between the heat source and the radiating surface is made either by direct contact or by means of heat pipes, see Sect. 13.2.6, although phase change capacitors are also used, whereas in the latter the thermal link is established by means of fluid loops, see Sect. 13.2.10, or by means of fluid loops plus heat pipes.

The temperature gradients along the radiators also play an important role in the heat that the radiator can exchange with space. These gradients are the result of the

Fig. 13.10 Sketch of a typical multilayer insulation



combination of heat conduction along the radiator and the actual heat radiation to space of each part of the radiator, leading to less than expected heat rejection values. Typically, this effect is measured by a figure of merit called the 'radiator efficiency', which is calculated as the ratio between the real heat radiated to the sink (considering the actual temperature distribution field along the radiator) and the heat that the radiator at constant maximum temperature would exchange with the sink. The main parameters that drive the radiator efficiency are thermal conductivity, radiator thickness, emissivity, and working temperature.

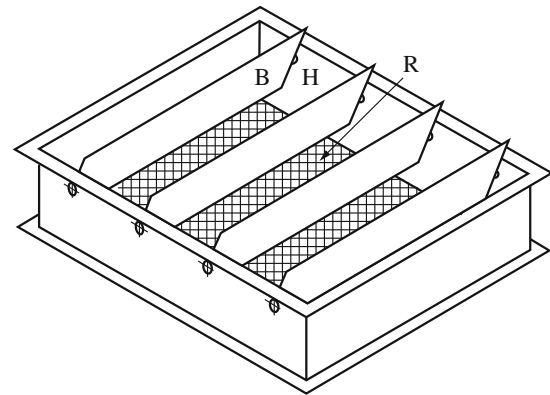


Fig. 13.11 Schematic diagram of a louver (Venetian blind type): *R* radiator; *H* housing; *B* blades

13.2.4 Louvers

Thermal louvers are active thermal control surfaces whose radiation characteristics can be varied in order to maintain the proper temperature of a component which experiences cyclical changes in the amount of heat that it absorbs or generates. This is normally achieved by means of blades whose orientation with respect to a fixed baseplate can be adjusted (see Fig. 13.11).

Louver systems can be made for shadow or sunlight operation. In the first case, heat is radiated through the louvers to the outer skin of the spacecraft, while in the second the excess heat is transferred from the emitting baseplate to outer space.

Louvers are composed of five main components: the baseplate, blades, actuators, sensing elements, and structural elements.

The baseplate is a surface of low absorptance-to-emissivity ratio which covers the critical set of components whose temperature is being controlled.

The blades, driven by the actuators, are the elements of the louvers which give variable radiation characteristics at the baseplate. When the blades are closed, they shield the baseplate from the surroundings, while when they are fully open, the coupling by radiation of the baseplate to the

surroundings is at its maximum. The radiation characteristics of the baseplate can be varied in the range defined by these two extreme positions of the blades.

The actuators are the elements of the louvers which drive the blades according to the temperature measured by sensors placed in the baseplate. Up to now, the actuators of the louvers used on satellites have been bimetal spirals or bellows, although other types could be used, such as Bourdon spirals, and electrical devices. In a single actuation system all the blades are driven by the same actuator. In the multiple blade actuator system several actuators are required to operate the system. Generally, bimetal is used in a multiple blade actuation system, and bellows in a single blade systems. Some recent missions like ESA's Rosetta spacecraft use also trimetallic coil spring actuators [21].

The sensing element senses the temperature of the baseplate and activates the actuators, which control the blades accordingly. The type of sensing element depends on the kind of actuator. When the actuator is a bimetal, the sensing element is the bimetal itself. If the actuators are bellows or Bourdon, the sensing element can be a tank or a tube containing either a liquid or a liquid-vapor mixture, and soldered to the baseplate.

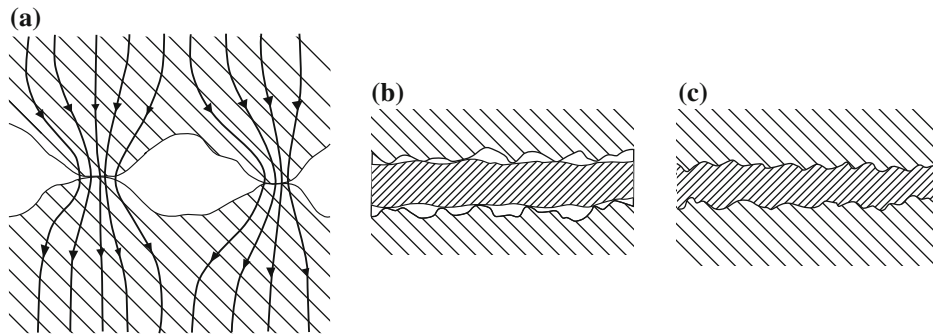


Fig. 13.12 **a** Schematic representation of two surfaces in contact with a heat flow across the interface; **b** Two surfaces with

uncompressed thermal filler; and **c** the surfaces with compressed thermal filler. In all cases the vertical scale has been exaggerated

13.2.5 Mechanical Interfaces

When two solids are brought into contact and heat is conducted from one to the other, a resistance to heat transfer appears in the interface caused by the inherent irregularities of the contacting surfaces. Each surface, no matter how well polished it is, consists of peaks and valleys as shown in Fig. 13.12. The actual solid-to-solid contact area is only a small fraction of the total apparent contact area. Voids formed by valleys are either empty in a vacuum environment or filled with gas in the presence of an atmosphere, the gas contributing little to the conduction of heat. The majority of the heat flow is constrained to the areas of solid-to-solid contact and gives rise to the observed temperature jump across the interface. This resistance to heat transfer is called the joint resistance or thermal contact resistance and the inverse value is the thermal contact conductance.

Numerically the thermal contact conductance is quantified as follows. Consider two solid rods A and B, of section A , with their ends kept at different temperatures T_{1A} and T_{2B} , as indicated in Fig. 13.13. Their lateral surfaces are thermally insulated so that heat conduction is one-dimensional. The heat flux is \dot{Q} . Macroscopically, in steady-state conditions, a temperature jump ΔT_c is observed in the contact plane. The thermal contact conductance h_c is defined as $h_c = \dot{Q}/(A\Delta T)$, and the thermal contact resistance is the inverse value $R_c = 1/h_c$. When the apparent area of contact cannot be easily identified (as for instance in a bolted joint), the thermal contact conductance is not defined per unit area, but simply $h_c = \dot{Q}/(\Delta T)$.

The thermal contact conductance is a fairly complex phenomenon, influenced by the contact pressure, the finish of the surface, and the mechanical properties of the contacting solids: the modulus of elasticity and the hardness. The design of the joints can be adapted to the thermal necessities of the system; if insulation is needed, bad couplings are sought, and if it is necessary to spread heat, good thermal couplings will have to be achieved.

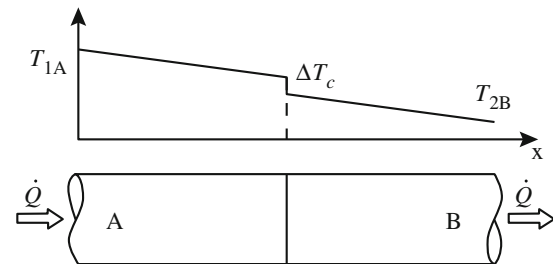


Fig. 13.13 Temperature distribution of two rods in contact, with the temperature jump at the interface, ΔT_c , indicated

A number of models are available in the literature to predict the thermal contact conductance between surfaces. Depending on the type of joint the models are organized in: bare or filled joints, in vacuum or in atmospheric environment, and in bolted or just pressed joints. A good compilation of the existing models can be found in [14].

Thermal fillers are materials used to fill the voids formed by the surface peaks and valleys appearing in the contact region when two materials are brought into contact, thereby enhancing the heat transfer. A sketch of how the interface is modified when a thermal filler is used is shown in Fig. 13.12. Thermal fillers are usually made of soft materials with high thermal conductivity, and in some cases also high electrical impedance so that they can provide electrical isolation. The main types of materials used as thermal fillers are graphite foils, elastomeric thermal fillers consisting of a thermoset elastomeric binder containing a dispersed highly thermally conductive ceramic filler, and room temperature vulcanizing (RTV) materials that also act as adhesive materials.

13.2.6 Heat Pipes

The heat pipe is a thermal device which allows an efficient transport of thermal energy. It consists of a closed structure containing a working fluid which transfers the thermal

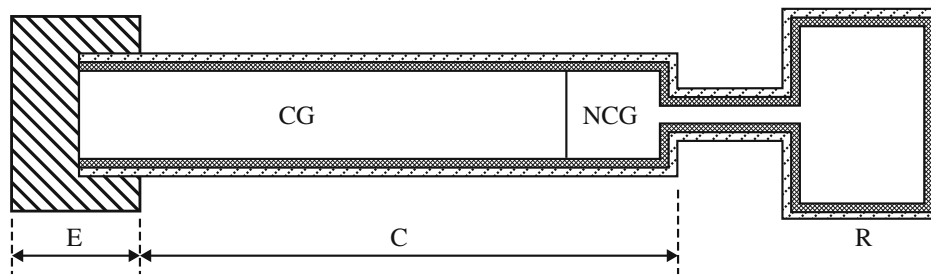


Fig. 13.14 Sketch of a variable conductance heat pipe: *E* evaporator; *C* condenser; *R* non-condensable gas reservoir; *CG* condensable gas; *NCG* non-condensable gas [7]

energy from one part (evaporator) to another (condenser), see Fig. 13.14. The phenomena involved in the transfer process are the following: (1) vaporization in the evaporator; (2) vapor flows in the core region of the container; (3) condensation in the condenser, and (4) liquid return to the evaporator by capillary action in the wick.

Because the pressure variations in the vapor core are normally small, the heat pipe temperature is nearly uniform and similar to the saturated vapor temperature corresponding to the vapor pressure. The capability to transport a large amount of thermal energy between two terminals (evaporator and condenser) with a small temperature difference is one of the main characteristics of heat pipes, which can be considered extra-high thermal conductivity devices in terms of Fourier's law. Their transport capacity is 4–5 orders of magnitude higher than that of a copper rod with the same mass.

Heat pipes can be made into different shapes, and by using the working fluid best suited to the desired temperature range, can operate at temperatures ranging from the cryogenic regions to high temperature levels which are only limited for structural reasons.

Pumping can be obtained by using wicks, grooved tubes, and arteries. The primary requirement for a heat pipe wick is that it should act as an effective capillary pump. That is, the surface tension forces developed between the fluid and the wick structure should be sufficient to overcome all viscous and other pressure drops in the pipe while maintaining the required fluid circulation.

Grooved tubes, with channels running axially along the inner surface of the tube, are structurally stable, have a large pipe wall to wick thermal conductance, and allow an easy control of pore size during manufacturing.

The condensed liquid can be delivered to the evaporator through one or more cylindrical arteries, which are placed near the pipe centerline, and are extensions of the wick covering the inner wall of the pipe.

A wide variety of fluids ranging from cryogenics to liquid metals have been used as heat pipe working fluids, like ammonia, ethanol, Freon 11 (trichlorofluoromethane, also

called CFC-11, or R-11), methanol, nitrogen, propane, water (saturated).

More details on heat pipes basics and applications can be found in [22].

13.2.7 Phase-Change Materials

Solid–liquid phase-change materials (PCM) present an attractive choice for passive thermal control of a spacecraft when the incident orbital heat fluxes or the on-board equipment heat dissipation fluctuate widely. Basically, the PCM thermal control device consists of a container that is filled with a substance capable of undergoing a phase change. When the temperature of the spacecraft surface increases, either because of external radiation or inner heat dissipation, the phase-change materials will absorb the excess heat through melting, and will restore it through solidification when the temperature decreases again.

To control the temperature of cyclically operating equipment, the phase-change material cell is normally sandwiched between the equipment and the heat sink.

Containers of PCM primarily differ in their outer shape (circular or rectangular), and in the flexible element which may be used for compensating the volume variation of the material during the phase change. The flexible elements of the container may be metallic bellows, metallic membranes, or rubber diaphragms.

When a rigid container is used it is necessary to provide a void or gas volume for expansion of PCM during melting.

The incorporation of thermal fillers into PCM systems offers distinct advantages, the primary one being the improvement of the thermal conductivity of the phase-change materials which, if not metallic, have low conductivities.

When fillers are not used, the temperature at the heated surface of the PCM may rise far above its melting point, with solid material still available but thermally isolated from the heated surface. However, when a metal filler is used, thermal gradients in the phase-change material bulk

are considerably reduced because of the high thermal conductivity of the filler.

A compilation of commercially available PCM systems and the materials they use can be found in [23].

13.2.8 Heaters

Reliable long-term performance of most spacecraft components takes place in a specified temperature range. The attainment of some temperature range requires, in many instances, the generation of heat within the spacecraft. In these cases, heaters are sometimes required to fulfill specific requirements like the protection of components for low temperatures, to provide precise temperature control for devices or components, or to warm up equipment to its operating temperature.

When a local uniform heat source or a profiled heating area is needed, electrical heaters can provide heat efficiently due to their versatility, although other types of heaters (chemical or nuclear) are also used in spacecraft. Obviously, the use of electrical heaters requires the availability of a power source. For near-Earth applications, solar power provided by photovoltaic devices is the preferred option because of the relative proximity of the Sun. However, when spacecraft missions must be performed far from the Sun or into harsh environments (such as the surface of Mars or in certain lunar locations), reliable, long-life electrical and thermal power sources independent of the Sun are needed (e.g. radioisotope heaters).

Electrical heaters are based on Ohm's and Joule's laws. Two types of heater typically used on spacecraft are film heaters (or patch heaters) and cartridge heaters. By far the most commonly used type is the film heater due to its flexibility, which means it can be installed on flat and curved surfaces. These are made of electrical resistance filaments sandwiched between two layers of electrically insulating material, such as Kapton, attached to leads.

In a radioisotope heater unit, the heat is produced from the natural decay of the radioisotope (alpha particles in the case of Plutonium-238). In this case, the waste heat from a radioisotope source is recovered by the spacecraft to provide additional thermal control for the avionics and instrumentation without resorting to additional electrical heaters. These heaters place the heat of radioactive decay directly where it is needed. Additional information on radioisotope heaters can be found in [24].

13.2.9 Heat Switches

Heat switches cannot really be classified as heaters, but their ability to adjust to variations in heat dissipation rates makes

them an attractive option for temperature control in modern satellites. If a heat switch connects an electronic component to a radiator, heat is removed from the device when it is generating large amounts of energy and conserved when the device is not producing heat, allowing the device to remain in the desired temperature range. Therefore, heat switches can passively control the temperature of warm electronics or instrumentation without the use of thermostats and heaters, thereby reducing power requirements.

In paraffin thermal switches, the volume change of paraffin, which expands approximately 15 % when it melts, facilitates heat switch operation. Under normal operating conditions, a paraffin heat switch contains a mixture of solid and liquid wax. In addition, a gap exists between the two devices connected by the paraffin heat switch. Due to the vacuum in the gap, heat transfer across the heat switch is limited to radiation across the gap and conduction through the support structure. When heat is added to the heat switch, it is absorbed as latent heat and melts some of the remaining solid paraffin. The melted paraffin expands and closes the gap that previously separated the hot and cold sides of the heat switch, enabling conduction across the entire heat switch surface. As more heat is added, more paraffin melts and the pressure at the contact between the hot and cold sides increases, causing an increase in conductivity.

Another type of thermal switch is the differential thermal expansion heat switch. This uses two materials with different coefficients of thermal expansion to control contact between the cold and hot sides of the switch. Additional details on thermal switches can be found in [25].

13.2.10 Fluid Loops

The aim of the fluid loop is to keep the temperature of a given component within the range that guarantees its correct functioning. To achieve this, the heat flow rate evacuated must be equal to the heat rate dissipated by the component, plus that coming from external sources.

In a fluid loop, the fluid is in motion, absorbing the heat at a relatively steady rate from the component whose temperature is to be controlled, and transferring it to a heat sink that can be separated from the source (Fig. 13.15). Heat transfer can be achieved through sensible heat change. A liquid or gas phase is used to transfer heat according to the equation $\dot{Q} = \dot{m}c_p\Delta T$. In order to increase the heat transfer rate, \dot{Q} , either the mass flow rate, \dot{m} , or the temperature difference, ΔT , has to be increased for a particular fluid (the specific heat, c_p , is more or less the same for all liquids and gases at normal conditions).

Forcing of the fluid through the duct can be performed by use of a pumping device (e.g. a centrifugal or positive displacement pump), normally driven by an electric motor.

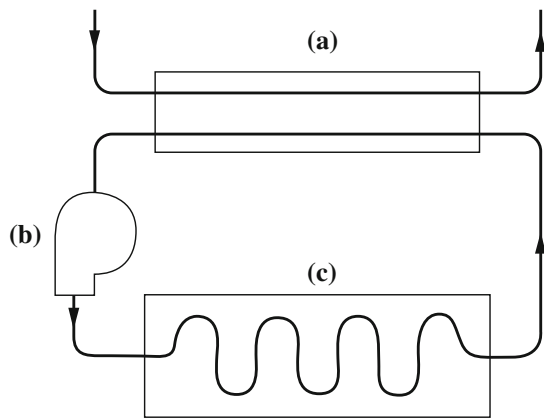


Fig. 13.15 Schematic representation of a fluid loop showing the main components: **a** heat exchanger, **b** pump, and **c** heat source

According to their working mode, fluid loops may be regarded either as thermal insulators (when the aim is to protect the component against a thermally hostile environment), or as thermal acceptors (transferring the excessive thermal energy from the component to the thermal sink).

The coolant may circulate through a single-walled chamber enclosing the component, or through a double-walled component heat exchanger package. The single-walled jacket affords more efficient heat transfer than the heat exchanger, but the fluid can contaminate, corrode or chemically react with the materials of the component that is being thermally controlled.

The heat rejection depends on whether the coolant is expendable or non-expendable. An expendable coolant is rejected from the vehicle once it has accomplished its mission, while a non-expendable coolant is recirculated after losing its excess thermal energy to space via a radiator.

When cryogenic cooling below 70 K is required, direct radiation to space is almost impossible. Because of this, the fluid loop incorporates a refrigerating system (Brayton, Stirling, or Vuilleumier cycles) which compresses the gas at an ambient temperature and then expands it at a lower temperature. During the expansion, heat is added to the gas providing the required cooling.

The fluid flow concepts and the performance of different fluid loop systems are extensively described in [7].

13.2.11 Thermoelectric Cooling

Thermoelectric coolers are solid-state devices that work as small heat pumps. They obey the laws of thermodynamics just as do conventional mechanical heat pumps (refrigerators), absorption refrigerators, and other devices involving the transfer of heat.

Using thermoelectric cooling in space applications has some advantages, mainly those regarding a flexible and

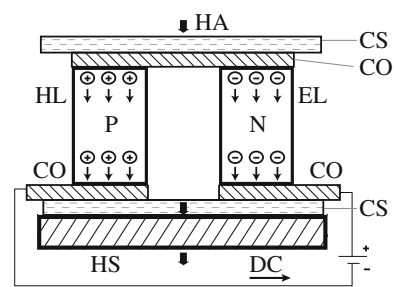


Fig. 13.16 Schematic of a thermoelectric cooling element: CO, copper plate; CS, ceramic substrate; DC direct current; EL electrons; HA heat absorbed; HL holes; HS heat sink; N, P, type N and type P semiconductors, respectively [26]

easy-to-control thermal regulation system. However, due to its low efficiency it is mostly suitable for localized cooling for temperature control of a single component, rather than a main cooling method for an entire system.

A thermoelectric cooler [26] consists of a type N and a type P semiconductor (such as of bismuth telluride) as shown in Fig. 13.16. A junction between these dissimilar semiconductors is formed at the surface to be cooled and a DC voltage is applied across the other junction at the hot surface where heat is transferred to the surroundings. The extra electrons in the N type material and the holes left in the P type material are the carriers that transfer the heat from the cold to the hot junction. The heat is pumped by virtue of the Peltier effect.

Thermoelectric elements are usually connected thermally in parallel and electrically in series to obtain the required power handling capacity.

13.2.12 Cryogenic Cooling

Systems that work under 100 K are called cryogenic systems. This temperature limit is typical of instruments devoted to Earth observation (infrared detectors) or gamma ray, high energy or infrared astronomy.

The most critical cryogenic cooling requirements in spacecraft subsystems come from instruments carrying detectors that have to work at very low temperatures. A detailed review of non-military space missions carrying cryogenic instruments is reported in [27, 28]. An overview of the cryogenic systems either for Earth or for space applications already developed or under development in Europe can be found in [29]. A comparative diagram of the applicability of several cryogenic cooling systems is shown in Fig. 13.17, outlining the region in the cooling power-temperature plane where these systems are placed.

Generally speaking, a cooling system provides a heat sink, evacuating the heat from the cool side of the equipment toward a hot part, where heat is dissipated. In the case

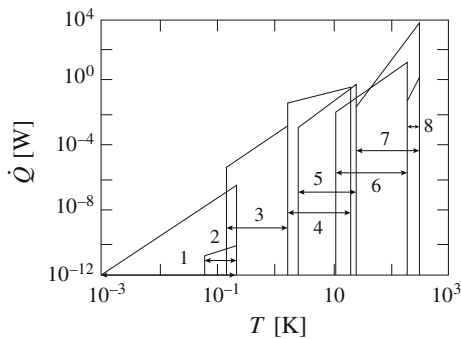


Fig. 13.17 Operational ranges of several cryogenic systems. Variation of the cooling power capacity, \dot{Q} , with temperature, T : 1 dilution/adiabatic demagnetization refrigerators; 2 solid state coolers based on normal metal–insulator–superconductor; 3 ^4He , ^3He sorption coolers; 4 liquid He, solid H_2 cryogenic tanks; 5 He, H_2 Joule–Thomson; 6 stirling, pulse tube; 7 radiators; 8 Peltier [27]

of a spacecraft, isolated in space, the evacuated energy is radiated to space either directly or by pumping energy to a high temperature level to radiate this energy more easily to space. In this second case, the heat pumping process can be performed following either a closed loop cycle or an open loop cycle (Table 13.4). Obviously, the final energy transference to space has to be performed using a radiator.

An open cycle involves the use of stored cryogenic materials, liquid or solids. In this case the work needed is performed on Earth, before the mission, in the liquefaction process (or solidification, as appropriate) of the working fluid. The cold heat sink is generated by the evaporation of cryogenic solids or liquids. In the open cycle systems, there is no heat radiation, although the gas generated by evaporation has to be evacuated. Thus, the system lifetime is driven by the heat losses and the cryogenic material stored on board.

In the closed cycle systems, mechanical coolers are employed, where work is continuously performed during operations.

13.2.13 Thermal Protection Systems

The thermal protection system (TPS) of a space vehicle ensures the structural integrity of its surface and maintains the correct internal temperatures (for crew, electronic equipment, etc.) when the vehicle is under the severe thermal loads of reentry. These loads are characterized by very large heat fluxes over the relatively short period of reentry. The heat fluxes acting on the TPS are so large because of the great speeds of reentry vehicles. For instance, the Space Shuttle velocity goes from approximately 8 km/s at an altitude of 100 km to 2 km/s at 50 km.

Generally speaking, the thermal protection system consists of a material system (shield and/or load carrying member) operating on a given heat dissipation principle.

Table 13.4 Cryogenic cooling systems

Radiators		
Open cycle	Solid cryogenes	
	Liquid cryogenes	
Closed cycle	Regenerative systems	Stirling
		Pulse tube
	Recuperative systems	Gifford
		Joule–Thomson
		Brayton

There are several thermal protection system concepts for reentry vehicles [30].

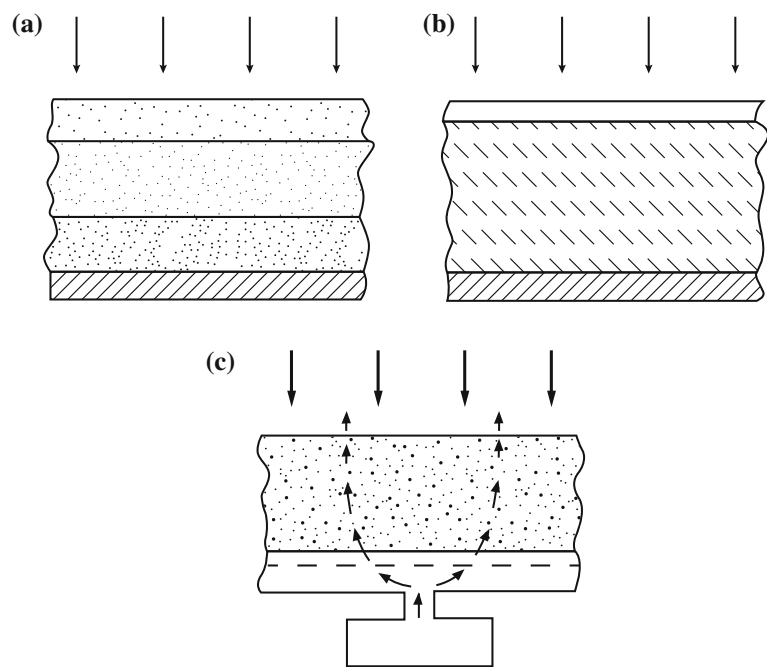
Ablative systems operate by dissipating the incident thermal energy through the loss of material. They have good thermal characteristics since phase changes absorb a large amount of energy. This concept has been widely used in most of non-reusable entry vehicles, for its simplicity and its high performance. It has been used in planetary probes [31], ballistic missiles and space capsules. See Fig. 13.18 for a sketch of an ablative system. Materials commonly used can be composites (carbon phenolic, silica phenolic, phenolic nylon), ceramics (graphite), metals (graphite) or plastics (silicone polymers, Teflon®—a brand name of polytetrafluoroethylene, AVCOAT or glass-filled epoxy-novolac).

Radiative systems operate by re-emitting the radiation energy received from the surrounding environment by the solid walls. They are composed of two layers: an outer layer, which consists of a refractory material that can stand the radiation equilibrium temperature, and an inner layer that insulates the outer layer from the structure in order to minimize the heat flow between the two, see Fig. 13.18. In these systems energy absorption is generally much smaller than for ablative systems.

Commonly used external insulators can be divided in two classes: rigid and flexible. Rigid insulators can adopt different shapes: tiles, shingles, shells, and boxes. They can be made out of composites (carbon/carbon, carbon/silicon carbide) or ceramics (sintered alumina/silica fibers, sintered high-purity silica fibers). This type of insulation is used to protect areas exposed to the highest temperatures. Flexible insulators are blankets of different materials: silica fiber, glass fiber, alumina/silica fiber, alumina/borosilicate fiber, Nomex® fiber, alumina fiber plated with rhodium, and nylon. These materials are processed into fleeces, felts or threads, which then form the blankets.

Transpiration systems are systems where fluid (H_2O , NH_3 , CF_4 , CO_2) is injected through a porous medium into the boundary layer [32]. The structure is maintained cool by exploiting two basic mechanisms. Firstly, heat is conducted to the coolant as it flows through the structure, and then as the coolant is ejected out the surface it reduces the surface

Fig. 13.18 **a** Sketch of an ablative thermal protection system; **b** a radiative thermal protection system; **c** a transpiration thermal protection system



heat transfer rate by cooling and thickening the boundary layer. See Fig. 13.18 for a sketch.

In some applications, the shape change caused by the surface recession of an ablating surface is not acceptable for, say, aerodynamic performance reasons. In such cases, if the environment is too severe for radiative or heat sink systems, transpiration cooling may be the only practical solution. This thermal protection system makes possible suitable performance in environments that could not otherwise be withstood. However, its mechanical complexity, with the associated reliability problems, tends to limit its use.

13.3 Thermal Control Design, Analysis and Testing

13.3.1 Thermal Control Design

The aim of a thermal control system (TCS) is to maintain all the components on-board the spacecraft within the allowable temperature limits by using the minimum spacecraft resources and controlling the heat fluxes through interfaces, as per equipment specifications. Furthermore, it has to guarantee the optimal performance of components in operational conditions. Besides keeping temperatures within ranges, the aim of the TCS is also to minimize temperature gradients according to specified limits and to guarantee temperature stability for optics, opto-mechanical devices, and any other components sensitive to temperature. This has to be done for all mission phases and the possible degradation that can be caused by the in-orbit environment (e.g. atomic

oxygen, ultraviolet radiation), wear and mechanical loads, has to be taken into account during the design process.

The two main tasks under the responsibility of the thermal control system team are: the definition of the thermal hardware of the spacecraft and the prediction of the temperatures achieved throughout the orbit, and the identification of the relevant parameters that have influence on the thermal behavior of the spacecraft in order to find an optimal solution compatible with the limitations given by the spacecraft resources.

The component requirements have to be defined for the different modes of operation of the spacecraft. They include operational mode, start-up, and survival conditions. In this last case, the goal of the thermal control system is to avoid damaging the equipment.

The thermal requirements in operational conditions of typical spacecraft equipment are listed in Table 13.5. Note that these are just typical values, shown as examples. For a given mission, the thermal requirements of the equipment and platform have to be specified because the requirements depend on the specific components used.

The thermal control system usually requires specific thermal hardware that has to be taken into account in the corresponding budgets. Minimum mass, power, and size have to be used as baseline criteria for this hardware definition, provided that reliability and safety requirements are fulfilled.

13.3.1.1 Design Process

The thermal control system design process consists of two main tasks. On the one hand, the appropriate thermal hardware for the spacecraft has to be selected. On the other,

Table 13.5 Typical temperature ranges for some spacecraft equipment

<i>Temperature ranges</i>	
Electronics (housing)	(−10 °C, +50 °C)
Batteries	(0 °C, +20 °C)
Solar arrays	(−100 °C, +120 °C)
Antenna dish	(−65 °C, +95 °C)
Hydrazine tank	(+10 °C, +50 °C)
Infrared detectors	(−223 °C, −173 °C)
Inactive structure	(−100 °C, +100 °C)
<i>Temperature gradients</i>	
Opto-electronic equipment	$\Delta T < 5$ °C
High resolution cameras	$\Delta T < 0.1$ °C
Detectors (CCD)	$\Delta T < 0.01$ °C
<i>Temperature stability</i>	
Electronics	$dT/dt < 5$ °C/h
Detectors (CCD)	$dT < 0.1$ °C during observation periods

the temperatures of the different parts of the spacecraft have to be calculated for different heat load cases, verifying that the thermal requirements are met.

There are many types of space missions and payloads, which means that the design of the spacecraft, and in particular of the thermal control system, has to be tailored for each type of mission. Hence, most communication satellites in geostationary orbits may be based on the same design philosophy whereas the mission requirements for low Earth orbit satellites or interplanetary spacecraft have a lot of influence on the system design.

The prediction of temperatures is obtained by solving the energy balance equation applied to the spacecraft. Obviously, the temperature distribution strongly depends on the thermal hardware used. Therefore, before carrying out any calculations, it is necessary to define an initial thermal hardware configuration of the spacecraft. This is commonly done based on engineering experience. For example, it is common practice to insulate the spacecraft from outer space using multilayer insulations (MLI). This helps to lessen the effect of the very variable environmental conditions on the equipment. In order to allow the rejection to space of the power that is internally dissipated, some radiators, located on the outer surface of the spacecraft subjected to lower environmental loads, facing deep space as much as possible, are appropriately sized. The thermal couplings between the inner equipment and the radiators are determined accordingly to enable the heat flux between dissipating devices and radiators. With the initial ‘guessed’ hardware, based on prior expertise, the temperatures of the spacecraft are determined and, depending on these results, the thermal hardware is modified until the requirements are met. The design process

is therefore an iterative process that has as its output the spacecraft thermal hardware configuration and the spacecraft’s temperature predictions. Furthermore, this iterative process involves not only the thermal control system but also other subsystems of the spacecraft. Indeed any change in the hardware may have direct implications in the mechanical and structural design, and the need for heaters has direct impact on the power management subsystem and the electronics and on-board data handling subsystems.

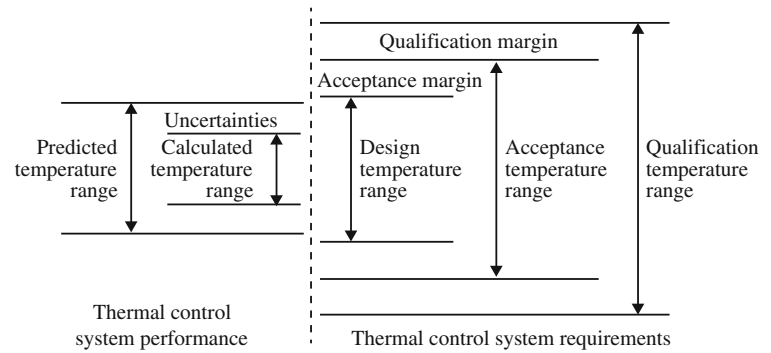
In the first loop of the iterative process, temperatures are calculated from semi-analytical simulations. This is always done in the early phases of a mission when the concept of the spacecraft is still not completely defined and detailed geometric information is still not available (phases 0, A and B, see Sect. 2.3.2: Fig. 2.9 and Table 2.2, and Sect. 7.2: Fig. 7.6 and Table 7.5). In these phases, it is common practice to work in parallel with more than one spacecraft configuration and apply trade-offs to select the final concept. However, due to the complexity of space systems and the capabilities of modern computers, very often simple numerical models are also used for trade-offs and parameter sensitivity analysis. Once the detailed definition phase starts (end of phase B and phase C), thermal calculations are always performed with complex numerical simulations that enable the determination of the spacecraft temperature field. Specific software tools, like for instance ESATAN or SINDA, are used for this purpose.

The inaccuracies of the design process (due to simplifications in the geometry, uncertainties in the properties of surfaces and materials, etc.) are also palliated by applying safety margins to the results predicted with the numerical models. Thus, the temperature range predicted with the models is enlarged with a margin that depends on the design phase and the level of detail of the models. In early phases of the design a typical uncertainty of ± 15 K is applied, but this margin may be reduced to ± 5 K after the mathematical models have been correlated with measured data obtained during the thermal balance tests. Figure 13.19 shows the philosophy of margins applied to the calculated temperature range in order to define the different levels of testing (qualification and acceptance tests), according to [33].

The major factors driving the thermal control system design are

- The spacecraft environment, which drives the external loads.
- The heat dissipated by the equipment on-board the spacecraft.
- The distribution of the thermal dissipation within the spacecraft.
- The temperature requirements of the spacecraft components.
- The configuration of the spacecraft: geometry, materials, mounting systems, etc.

Fig. 13.19 Temperature margins definition for thermal control subsystem



13.3.1.2 Load Cases

Spacecraft are subjected to highly variable environmental conditions (see Sect. 13.1.3). The thermal control system has to fulfill thermal requirements over all mission phases. This includes ground operations (integration, testing, storage, and shipping) and flight activities (launch, transfer orbits, cruise, final orbit, etc.). In order to size the thermal control system, the worst case scenarios, those leading to the extreme thermal loads, have to be identified. Once identified, the so-called hot cases and cold cases are defined for operating, start-up, and survival conditions of equipment. These dimensioning cases are defined by an appropriate combination of external fluxes (solar, albedo, and planetary infrared), material properties, and unit dissipation profiles. Normally, the hot case corresponds to the maximum external loads and maximum internal dissipation. The maximum external loads usually occur at the sub-solar point in a planetary orbit or at perihelion in a solar orbit. The cold case usually corresponds to eclipse zones for planetary orbits and to aphelion for a solar orbit. Modes of operation with minimum dissipation are chosen to assess the cold cases.

A common philosophy in spacecraft thermal control is to design the thermal subsystem for the hot operational case. However, as the radiator sizes selected for the hot case could lead to extremely low temperatures when the spacecraft is exposed to cold conditions or the equipment is off, electrical substitution heaters are appropriately sized and located for these cold situations.

The degradation of surface properties, for instance the increase of solar absorptance values, α , with prolonged exposure to solar radiation, also has to be taken into account. End of life (EOL) property values have to be used for thermal calculations corresponding to the hot cases, whereas beginning of life (BOL) property values have to be used to define the cold cases.

Since the dimensioning loads are defined for the worst case scenarios, steady-state calculations under these conditions are carried out. This is a conservative approach that simplifies the calculations and provides the upper and lower limits that temperatures can reach.

Regarding the different mission phases, the thermal control system is usually designed for the cruise and nominal orbit, making it compatible with ground operations and the ascent phase.

13.3.2 Thermal Mathematical Models

Thermal modeling is the major task in thermal control system design.

The thermal modeling process requires approximating the physical system by a mathematical representation, that is, a set of numbers that represent the system from the thermal point of view. The first step in the mathematical procedure is to set up the so-called geometrical mathematical model (GMM). This representation of the geometry of the system is necessary to compute the external thermal loads on the outer surfaces and the radiation exchange between different parts of the spacecraft. It is usually a simplified geometry, where irrelevant details from the thermal point of view are excluded.

Once a basic geometry for thermal analysis is defined, this geometry is discretized in a network of nodes. To do this, the basic geometric shapes that comprises the geometrical mathematical model are meshed. Each node is an isothermal element characterized by its temperature, T_i , and its thermal capacitance, C_i . This numerical approach is called a lumped parameter network because the continuous parameters of the thermal system have been lumped into the discrete set of nodes.

The energy equation for each node can be written as

$$C_i \frac{dT_i}{dt} = \dot{Q}_{sun,i} + \dot{Q}_{alb,i} + \dot{Q}_{planet,i} + \dot{Q}_{dis,i} + \sum_{j=1}^n K_{ij}(T_j - T_i) + \sum_{j=0}^n R_{ij}(T_j^4 - T_i^4) \quad (13.30)$$

where, $\dot{Q}_{sun,i}$, $\dot{Q}_{alb,i}$, $\dot{Q}_{planet,i}$ are the external thermal loads at node i for solar, albedo, and planetary infrared respectively, $\dot{Q}_{dis,i}$ is the power dissipated at node i , and K_{ij} and R_{ij} are the

conductive and radiative links or couplings between nodes, respectively. Note that $K_{ij} = K_{ji}$ and $R_{ij} = R_{ji}$. Thus, the term $\sum_{j=1}^n K_{ij}(T_j - T_i)$ represents the conduction heat received by node i from the rest of the neighboring nodes j , and $\sum_{j=0}^n R_{ij}(T_j^4 - T_i^4)$ represents the net radiation exchange on node i .

Note that in the last term of Eq. 13.30, node 0 represents outer space. Thus, a radiative link between the spacecraft and outer space is taken into account whereas a conductive link is not considered.

When Eq. 13.30 is applied to all the nodes that the spacecraft is divided into, a system of ordinary differential equations is obtained and its solution allows the temperature of such discrete points to be determined. This means that in order to determine the temperature of the satellite, two matrices of coefficients (conductive and radiative thermal couplings) and four vectors (solar, albedo, Earth infrared, and internal dissipation thermal loads) of Eq. 13.30 are necessary. These matrices and vectors constitute a mathematical representation of the thermal model of the spacecraft by concentrated thermal capacitance nodes, coupled by a network made of thermal conductors (mainly radiative and conductive). That is why that set of numbers is called thermal mathematical model (TMM).

For thermal control system dimensioning, as said above, steady-state calculations for the worst-case hot and cold scenarios are performed. This is done by setting the left hand side of Eq. 13.30 to zero. When applied to all the nodes, this constitutes a system of algebraic non-linear equations that can be solved to obtain the temperature distribution for such extreme cases. In fact, what is solved in this case are simple heat flux balances. It is obvious that prior to the resolution of the equation it is necessary to obtain the thermal loads and the matrices of thermal couplings. These temperatures are the upper and lower limits that may be encountered during the mission. It is important to point out that the thermal design is usually obtained from these steady calculations whenever there are no stringent stability requirements.

Once a feasible thermal control system design based on the previous steady-state calculations is found, transient analyses are performed to determine the changes of temperatures over time and to verify the fulfillment of stability requirements. In this case, the system of equations is generally integrated using the Crank–Nicholson method.

13.3.3 Thermal Control Testing

Tests are needed as part of the verification process of the spacecraft thermal control system, together with the analysis with the mathematical models that have been described previously.

Thermal control testing covers different objectives

- To confirm that the system will operate satisfactorily at expected (or more extreme) operating temperatures.
- To evaluate the ability of the thermal control system to maintain the spacecraft thermal environment within established structural, experimental, and subsystem temperature limits.
- To verify the validity of the mathematical model.

Different types of test are required to accomplish these objectives. The first one is achieved by means of a thermal vacuum test, and the other two by the so-called thermal balance test. According to [33], conformance to specified performance has to be demonstrated by performing thermal balance, thermal vacuum, and climatic tests at all temperature ranges.

The test levels in thermal balance testing are set to simulate the external environment (solar radiation and deep space) or to approximate the anticipated energy flux levels at the boundaries of the spacecraft. These levels are then used in the mathematical model in order to permit valid comparison with the test. For thermal vacuum testing, temperatures are set equal to or higher than expected flight temperatures by a given margin. For both types of testing, the component electrical dissipation rates and duty cycles are set to values appropriate to the mission mode being tested. In some cases, it may be both technically and economically advantageous to perform a combined thermal balance and thermal vacuum test. Test conditions have to be agreed with the system authority and included in the system test plan.

13.3.3.1 Model Philosophy (Structural Thermal Model, Qualification Model, Flight Model)

Thermal testing is performed at various stages of spacecraft development according to the needs of the particular program.

The verification by testing (see Chap. 7) is implemented on the selected models chosen for the project or the model philosophy adopted. The model philosophy is defined by means of an iterative process that combines programmatic constraints, verification strategies, and the integration and test programs, taking into account the development status of the candidate design solution.

Generally, the qualification, acceptance and proto-flight test activities are distributed among the different models. This sharing depends on the model philosophy, the project characteristics, and the model representativeness.

The hardware models related to the verification of the thermal control system are

- The thermal model (or structural-thermal model STM)
- The qualification model (QM)
- The flight model (FM)
- The proto-flight model (PFM).

For instance, the main purpose of qualification thermal vacuum testing performed on either the structural-thermal model (STM) or qualification model (QM) is to detect the adverse effects on spacecraft performance that could result from any existing weakness in the thermal design. The purpose of acceptance thermal vacuum testing is to reveal adverse effects arising from defects in materials or workmanship in the flight model related to the thermal design. Both types of testing involve the collection and analysis of spacecraft performance data; the role of the thermal control engineer is to ensure that the spacecraft is exposed to the specified environments.

13.3.3.2 Development Tests

Thermal testing is addressed in [33] at the component and system (space vehicle) level. Besides these two levels, special tests may be necessary at other assembly levels. In addition, dedicated tests may be required to provide confidence in a new design or to aid the analysis. Development tests can provide early data to assist in the design or manufacturing process. For instance, thermal cycling tests are used to demonstrate the ability of equipment to fulfill all functional and performance requirements over the qualification temperature range at ambient pressure.

13.3.3.3 Thermal Vacuum Tests

The thermal vacuum testing is performed at the component, subsystem, and integrated spacecraft levels. At the component level, the testing is generally performed at the manufacturer's facility to ensure that the unit meets its reliability and quality assurance requirements.

At the subsystem level, there are design qualification and flight acceptance tests. The purpose of the design qualification test is to prove the component design by checking its performance capability in a vacuum under temperature stress more severe than predicted for the mission. A prototype component is generally used for design qualification testing. The flight acceptance test is performed on a flight model component and its purpose is to locate latent material and workmanship defects in a component of proven design by checking its performance capability in a vacuum at the temperature extremes expected in flight.

The purpose of acceptance testing on the flight model of the spacecraft is to check the interaction between subsystems as well as to ascertain the proper operation of all systems.

Test durations for acceptance thermal vacuum testing must be long enough to demonstrate that the unit can survive the launch and flight. Test times for qualification testing are not as easily defined, because testing is not performed on a flight unit and the test levels are more severe than encountered in flight.

13.3.3.4 Thermal Balance Tests

Thermal balance testing for design or development is performed to provide design information on those components for which the thermal design is difficult to analyze, stringent temperature constraints are imposed, or it is necessary to establish the feasibility of the design approach. Thermal balance testing is also conducted for design verification.

According to [33], for thermal control system items controlled by radiative and conductive heat exchange, a thermal balance test has to be performed in order to

- Provide data for the verification of the thermal mathematical model as part of the thermal control system qualification.
- Demonstrate the suitability of the thermal control system design.
- Verify the performance of thermal control system hardware
- Provide data about the sensitivity of the thermal control system design to parameter changes (for example, heat dissipation).

Thermal balance testing is generally performed on items at high integration levels, such as spacecraft, service module, payload module, or instruments.

The test instrumentation and the test set-up to be used must be defined in the test specification and agreed with the system authority; for example, temperature sensors and heaters with adequate number and position.

The thermal balance test has to provide accurate and reliable input data for the thermal model correlation [33]. Two different steady-state test cases have to be performed. A transient case has to be included for items that are sensitive to dynamic behavior.

The duration of thermal balance testing can be determined in two ways: (1) the test conditions are established and held until the test article reaches temperature stabilization, and (2) the test conditions are varied to simulate transient conditions in the same time frame as expected in flight.

13.3.3.5 Test Facilities

Thermal tests are performed in vacuum chambers where the temperature and heat flux can be controlled. There are several different methods for simulating flux in thermal balance testing. Heating can be provided by means of electrical heaters or halogen lamps. For cooling, liquid or gaseous nitrogen are normally used to achieve the low temperatures found in space.

Temperature conditioning in thermal vacuum testing is usually accomplished by varying the test chamber wall temperature or by monitoring the test article on a temperature controlled baseplate. The test specimen is heated by radiation when the purpose is to simulate space conditions, normally for full spacecraft tests or for equipment located at

the outer part of the vehicle (antenna reflectors, solar panels, etc.). Test specimens representing equipment located inside the satellite are heated by conduction through the baseplate on which the equipment is mounted.

Tests are often performed without regard to possible errors in the test chamber. The consequent test results could dictate an unnecessary redesign or confirm thermal adequacy of a deficient design.

There are several sources of error associated to the test facility and test set-up

- Conduction transfer from the fixtures used in mounting and supporting test articles in the chamber.
- Infrared energy inputs to test articles from the chamber and reflection from chamber walls and fixtures.
- Monitoring errors (calibration and measurement).
- Thermal losses to wiring.
- Deviation from the programmed cycle, or simulation errors.

Acknowledgments Except Figs. 13.15 and 13.18, all the figures of this chapter are from [15] and the authors are indebted to Woodhead Publishing Ltd. for their permission.

References

1. Çengel, Y.A., "Heat and Mass Transfer: A Practical Approach," McGraw Hill, New York, 2007.
2. Incropera, F.P., DeWitt, D.P., Bergman, L.T., and Lavine, A.S., "Fundamentals of Heat and Mass Transfer," John Wiley & Sons, New York, 2007.
3. Holman, J., "Heat Transfer," McGraw Hill, New York, 2010.
4. Modest, M.F., "Radiative Heat Transfer," Academic Press, Amsterdam, 2003.
5. Howel, J.R., Siegel, R., and Menguc, M.P., "Thermal Radiation Heat Transfer," CRC Press, Boca Raton, 2011.
6. "Data for selection of space materials and processes", ECSS-Q-70-71A rev. 1. June 2004.
7. "Spacecraft Thermal Control Design Data Handbook", ESA PSS-03-108, Issue 1, 1989.
8. "Terrestrial Environment (Climatic) Criteria Handbook for Use in Aerospace Vehicle Development," NASA-HDBK-1001, August 2000.
9. "Space Engineering - Space Environment", ECSS-E-ST-10-04C, November 2008.
10. "Space environment (natural and artificial). Process for determining solar irradiances," ISO 21348, May 2007.
11. "Space Engineering. Mechanical – Part 1: Thermal Control," ECSS-E-30 Part 1A, April 2000.
12. "Standard Solar Constant and Zero Air Mass Solar Spectral Irradiance Tables," ASTM E490-00a, April 2006.
13. Anderson, B.J., Justus, C.G., and Batts, G.W., "Guidelines for the Selection of Near-Earth Thermal Environment Parameters for Spacecraft Design," NASA TM 2001/211221, October 2001.
14. Gilmore, D.G., ed., "Spacecraft Thermal Control Handbook, Vol. I: Fundamental Technologies," 2nd edn., The Aerospace Press, El Segundo, 2002.
15. Meseguer, J., Pérez-Grande, I., and Sanz-Andrés, A., "Spacecraft Thermal Control," Woodhead Publishing, Oxford, 2011.
16. Touloukian, Y.S., DeWitt, D.P., and Hernicz, R.S., "Thermal Radiative Properties. Coatings," *Thermophysical Properties of Matter*, Vol. 9, IFI/Plenum, New York, 1972.
17. Henninger, J.H., "Solar Absorptance and Thermal Emittance of Some Common Spacecraft Thermal-Control Coatings," NASA RP-1121, April 1984.
18. Kauder, L., "Spacecraft Thermal Control Coatings References" NASA/TP-2005-212792, December 2005
19. Sanz-Andrés, A., Santiago-Prowald, J., and Ayuso-Barea, A., "Spacecraft Launch Depressurization Loads," *Journal of Spacecraft and Rockets*, Vol. 34, No. 6, 1997, pp. 805-810. doi: [10.2514/2.3290](https://doi.org/10.2514/2.3290)
20. Finckenor, M.M., and Dooling, D., "Multilayer Insulation Material Guidelines", NASA/TP-1999-209263, April 2009.
21. Domingo, M., and Ramirez, J.J., "Mechanical design and test of ROSETTA Platform Louvres," *Proceedings of the 10th European Space Mechanisms and Tribology Symposium*, ESA SP-524, September 2003, pp. 289-292.
22. Silverstein, C., "Design and Technology of Heat Pipes for Cooling and Heat Exchange," Taylor & Francis, Washington, 1992.
23. Zalba, B., Marin, J.M., Cabeza, L.F., and Mehling, H., "Review on thermal energy storage with phase change: materials, heat transfer analysis and applications," *Applied Thermal Engineering* Vol. 23, 2003, pp. 251-283.
24. Rinehart, G.H., "Design characteristics and fabrication of radioisotope heat sources for space missions," *Progress in Nuclear Energy*, Vol. 39, 2001, pp. 305-319.
25. Hengeveld, D.W., Mathison, M.M., Braun, J.E., Groll, E.A., and Williams, A.D., "Review of Modern Spacecraft Thermal Control Technologies," *HVAC&R Research*, Vol. 16, 2010, pp. 189-220.
26. Scott, A.W., "Cooling of Electronic Equipment," John Wiley & Sons, New York, 1974, Chap. 8, pp. 215-227.
27. Collaudin, B., and Rando, N., "Cryogenics in space: a review of the missions and of the technologies," *Cryogenics*, Vol. 40, 2000, pp. 797-819.
28. Donabedian, M., ed., "Spacecraft Thermal Control Handbook, Vol. II, Cryogenics," The Aerospace Press, El Segundo, 2003.
29. Ravex, A., and Trollier, T., "Recent developments on cryocoolers in Europe," *Proceedings of the Twentieth International Cryogenic Engineering Conference*, ICEC 20, Beijing, 2005, pp. 127-136.
30. Laub, B., and Venkatapathy, E., "Thermal Protection System technology and facility needs for demanding future planetary mission," *Proceedings of the International Workshop on Planetary Probe Atmospheric Entry and Descent Trajectory and Science (IPPW1)*, ESA SP-544, February 2004, pp. 239-247.
31. Poncy, J., Lebleu, D., Arfi, P., and Schipper, A.M., "Entry descent and landing systems for future missions," *Acta Astronautica*, Vol. 67, 2010, pp. 173-179.
32. Shuang, L., and BoMing, Z., "Experimental study on a transpiration cooling thermal protection system," *Science China*, Vol. 53, 2010, pp. 2765-2771.
33. "Space engineering. Thermal control general requirements", ECSS E-ST-31C, November 2008

Further Reading

34. Brown, C.D., "Elements of Spacecraft Design," AIAA Educational Series, AIAA, Reston, 2002.
35. ESA "Space Engineering. Mechanical – Part 1: Thermal Control," ECSS-E-30 Part 1A, April 2000.
36. ESA "Spacecraft Thermal Control Design Data Handbook", ESA PSS-03-108, Issue 1, 1989.

37. Fortescue, P., Stark, J., and Swinerd, G., eds., "Spacecraft Systems Engineering," 3rd edn., Wiley, Chichester, 2003.
38. Gilmore, D.G., ed., "Spacecraft Thermal Control Handbook, Vol. I: Fundamental Technologies," 2nd edn., The Aerospace Press, El Segundo, 2002.
39. Karam, R.D., "Satellite Thermal Control for System Engineers," Progress in Aeronautics and Astronautics, Vol. 181, AIAA, Reston, 1998.
40. Meseguer, J., Pérez-Grande, I., and Sanz-Andrés, A., "Spacecraft Thermal Control," Woodhead Publishing, Oxford, 2011.
41. Pisacane, V.L., "Fundamentals of space systems," Oxford University Press, Oxford, 2005.
42. Tribble, A.C., "The Space Environment, Implications for Spacecraft Design," Princeton University Press, Princeton, 2003.

Ali Atia and Huiwen Yao

Communications satellites are signal relay stations in orbit around the Earth. A satellite communications system, as shown in Fig. 14.1, includes a ground segment and a space segment. The ground segment consists of Earth stations/terminals which provide direct communications, including telemetry, tracking, and command (TT&C), to the space segment and network control center(s), which provide the network management and traffic control. A space segment is one or more spacecraft on-orbit which provides one or all of the following functions: (a) communications relay; (b) communications signal processing and traffic switching/redirection; and (c) data collection and transmission to ground stations. See Chap. 2 for a more detailed discussion of ground and space segments.

While there are many satellites in various Earth orbits, the most common systems use satellites in geostationary Earth orbit (GSO or GEO). As detailed in Sect. 4.4.3, this unique orbit is in the equatorial plane at an altitude of approximately 36,000 km, where the orbit period is equal to Earth's sidereal rotation period of approximately 24 h. As such, GEO satellites appear stationary when viewed by a ground observer.

The concept of the geostationary communications satellite was apparently first introduced by the Austro-Hungarian rocket engineer and pioneer of cosmonautics Herman Potočnik (pseudonym Hermann Noordung; 1892–1929) in his book *Das Problem der Befahrung des Weltraums—der Raketen-Motor* (The problem of space travel—the rocket motor) [1]. It should however be noted that it was, most likely, Konstantin Tsiolkovsky (1857–1935), who first conceived of the concept of a geostationary orbit. Potočnik first recognized the advantages of a spacecraft in such an orbit for communications and Earth observation. Later, in

October 1945, Arthur C. Clarke (1917–2008), published an article titled *Extra-terrestrial Relays* in the British magazine *Wireless World* [2]. The article described the fundamentals behind the deployment of artificial satellites in geostationary orbit for the purpose of relaying radio signals and, in effect, introduced the concept into the Western literature. Although Clarke is often credited with being the inventor of the communications satellite, his true influence is unclear [3].

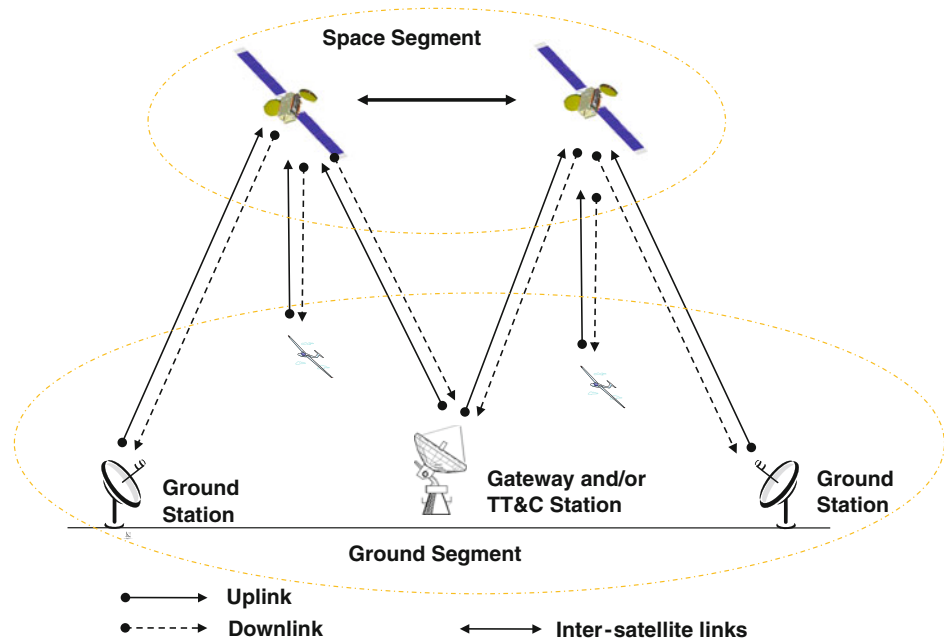
Telstar was the first active, direct relay communications satellite; see Fig. 1.7. It was launched by NASA from Cape Canaveral on July 10, 1962, as part of a multinational agreement between AT&T, Bell Telephone Laboratories, NASA, the British General Post Office, and the French National PTT (Postes, télégraphes et téléphones) to develop satellite communications. Telstar was placed in an elliptical orbit with an apogee of 6,000 km and a perigee of 1,000 km, giving an orbit period of 2 h and 37 min, in a plane inclined at 45° to the equatorial plane.

Hughes' Syncom-2, launched on July 26, 1963, revolved around the Earth once per day at a constant speed, but because its orbital plane was inclined to the equator it had a north–south motion as seen from the ground, and special equipment was needed to track it. The first truly geostationary satellite was Syncom-3, launched on August 19, 1964. It was placed in orbit at 180° east longitude and was used to relay experimental television coverage of the 1964 Summer Olympics from Tokyo, Japan, to the United States, making these Olympic Games the first to be broadcast internationally.

On August 20, 1964 The International Telecommunications Satellite Consortium (INTELSAT) was established on the basis of agreements signed by governments and operating entities, with the goal of establishing a global satellite system. Shortly after, Intelsat-1, also known as Early Bird, was launched on April 6, 1965 and placed in orbit at 28° west longitude. It was the first commercial geostationary satellite for telecommunications over the Atlantic Ocean,

A. Atia · H. Yao (✉)
Orbital Sciences Corporation, Dulles, VA, USA
e-mail: Yao.Huiwen@orbital.com

Fig. 14.1 A satellite communications system consisting of a ground segment and a space segment



and ‘live via satellite’ was born. On July 1, 1969 the world’s first global satellite communications system was complete with the Intelsat-3 satellite covering the Indian Ocean region. The 1970s was a decade of expansion in which commercial global communications via the Intelsat system greatly expanded, and several domestic satellite systems (Weststar, RCA-Satcom and Anik for Canada, Telstar and Comstar in the US, and Palapa in Indonesia) were established. The early era of the 1970s was all C-band (6/4 GHz) satellite systems. The experimental communications technology satellite Communications Technology Satellite, also known as ‘Hermes’ launched in 1976 demonstrated the commercial viability of the 14/12 GHz Ku-band for direct-to-home (DTH) satellite broadcasting. This gave rise to an upsurge of the 14/12 GHz satellites in the 1980s and 1990s. The Ku-band frequency continues to be the most sought after by service providers. The 1980s and 1990s were decades where commercial satellite communications and DTH broadcasting by satellites expanded through the introduction of more advanced technologies that resulted in more powerful satellites and more efficient use of the available frequency spectrum. Deregulation of telecommunications worldwide resulted in the establishment of many national and regional satellite systems. The 21st century is witnessing the evolution of satellite communications into an essential part of everyday life: the Internet, mobile communications, increasingly high definition television broadcasting, etc. Satellites became much more powerful with the introduction of many new technologies that enabled the use of more frequency bands and orders of magnitude of increased communications capacity and throughput.

Major commercial satellites manufacturers include Astrium, Boeing Satellite Systems (formerly Hughes), Lockheed Martin Space Systems, Orbital Sciences Corporation, Space Systems/Loral, and Thales Alenia Space. Other manufacturers vying for the commercial market are China Great Wall Industries, ISRO of India, and Melco of Japan.

14.1 Frequency Spectrum and Bands Allocations

The frequencies used for satellite communications are mainly determined by three factors

1. *Absorption by the atmosphere as a function of frequency*—The average atmospheric absorption varies as a function of frequency at different altitudes above sea level, along with the effects of rain and fog. The absorption has peaks due to different molecules in the atmosphere at particular frequencies. Usually these frequencies are avoided for communications applications, though in special cases they may be deliberately used so that the signal will not propagate beyond a certain range—e.g. covert military signals, or mobile communications where the limited frequency range available means that the same frequency may be reused many times in different communications cells. Frequency bands for satellite communications were allocated to lie within windows of least atmospheric absorption.
2. *International agreements/regulations*—The use of different frequency bands for different applications has been agreed through various international agencies.

Table 14.1 Radio frequency bands for space communications

Frequency band designation	Uplink frequency bands	Downlink frequency bands	Application
UHF	225–460 MHz	225–400 MHz	Military, Mobile Satellite Service (MSS)
L-Band	1,610–1,660 MHz	1,525–1,559 MHz	MSS
S-Band	2.65–2.69 GHz	2.48–2.65 GHz	Fixed Satellite Service (FSS), MSS, research
C-Band (including extended C-Band)	5.85–6.7 GHz	3.4–4.8 GHz	FSS
X-Band	7.9–8.4 GHz	7.25–7.75 GHz	FSS, military communication, and Earth observation satellites
K and Ku-Band	12.75–14.8 GHz	10.7–12.7 GHz	FSS, Broadcast Satellite service (BSS)
	17.3–18.4 GHz	17.7–18.4 GHz	
Ka-Band	27–31 GHz	18.4–22.0 GHz	FSS, MSS, research, and Intersatellite links
		31.8–32.3 GHz	
Q-Band	40.5–43.5 GHz	37–42.5 GHz	BSS, FSS, BSS, and research
V-Band	46–56 GHz		BSS, MSS, Intersatellite links
W-Band	56–100 GHz		Intersatellite links, FSS, and MSS

The International Telecommunications Union (ITU) is the specialized agency of the United Nations that is responsible for information and communication technologies. ITU coordinates the shared global use of the radio spectrum, promotes international cooperation in assigning satellite orbits, works to improve telecommunication infrastructure in the developing world, and establishes worldwide standards. It also publishes the Radio Frequency Regulations, which contain detailed allocations of frequency bands for all telecommunications applications, both terrestrial and satellite-based.

3. *The antenna size needed to produce a beam with the required angular spread*—The basic (approximate) relationship between wavelength and antenna size is θ (radians) $\approx \lambda/D$ where θ is the angular breadth of the main beam between the 3 dB points, which is often referred as beamwidth, and D is the maximum dimension across the antenna aperture. A satellite's antenna size must be chosen to produce the required coverage, and to fit inside the launch vehicle. An Earth station's antenna size must be chosen to produce the narrowest possible beamwidth capable of providing an adequate gain for communicating with the desired satellite and to avoid unwanted interference to and/or from satellites that may be in nearby orbit locations. For commercial applications, such as DTH, the antenna size must be as small as possible for economic reasons. At low frequencies, the wavelength is large and a large antenna is necessary to avoid interference. As the frequency increases, the beamwidth reduces for a given antenna size but the attenuation of the atmosphere increases, suggesting that the antenna gain should be higher, possibly necessitating a larger aperture. A

compromise must be made. Note that atmospheric attenuation is not a problem for satellite-to-satellite links, so these may involve millimeter-wave frequencies and very small antennas.

Table 14.1 provides information about the most common designations for the satellite frequency bands. The radio frequency allocations in the US can be found in a comprehensive chart at the NTIA web site [4].

14.2 Communications Systems Overview

There are several ways to classify satellite communications system architectures. The specific architecture is chosen according to the intended satellite communications services. These services may include two-way voice, two-way balanced data, services associated with very small aperture terminals (VSAT), video distribution to cable head-ends, DTH broadcasting, multimedia Internet access (two way unbalanced data), and mobile communications [5].

The basic satellite architectures are designated as non-processing (bent-pipe) or processing. A bent-pipe satellite simply takes the uplink signal from Earth, converts it to a downlink frequency, and amplifies it for transmission back to the Earth, possibly cross-connecting subchannels to other downlink antenna beams. A processing satellite may actually separate and demodulate the uplink signals and route the baseband signals to other beams, upconvert them to the downlink frequency band and send them down on the appropriate beams. Both processing and non-processing satellite architectures include antennas for receiving and

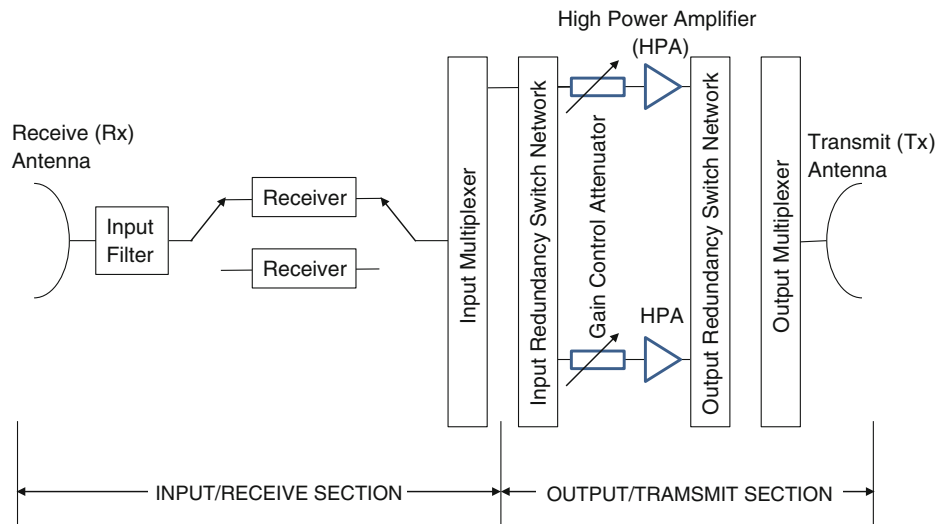


Fig. 14.2 Simplified functional block diagram of a bent-pipe communications payload

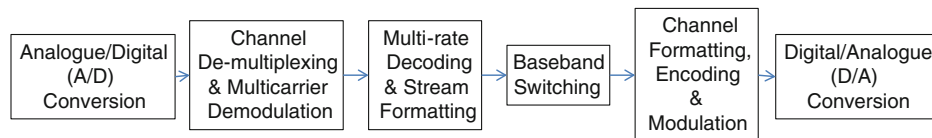


Fig. 14.3 Architectural block of a regenerative processor

transmitting in either broad spatially shaped beams, covering large regions of the Earth (for example, the continental United States or Europe), or in narrow spot beams to cover smaller regions.

Most GEO communications satellites employ the non-processing, channelized communications system architecture. In this architecture, the satellite receives signals from a ground station, amplifies the signals, changes their carrier frequencies, subdivides the received spectrum into several channels (transponders), power amplifies each of these channels, combines (multiplexes) the amplified channels and retransmits the combined signals to the receiving ground stations. Figure 14.2 shows a typical bent-pipe non-processing payload.

Processing payload architectures have been developed and used for specialized applications ranging from mobile narrowband connectivity [4], to wideband Internet/data connections, to Digital Video Broadcasting by Satellites (DVB-S) [6]. Processing payloads can be regenerative or transparent (non-regenerative). The regenerative on-board processors demodulate the signals and manipulate the baseband bits to perform baseband switching and routing. A typical architectural block diagram of a regenerative on-board processor's functions is shown in Fig. 14.3. Transparent processing payload architectures are used to digitally route, interconnect, and switch narrowband analog channels from a large number of uplink beams to a large number of

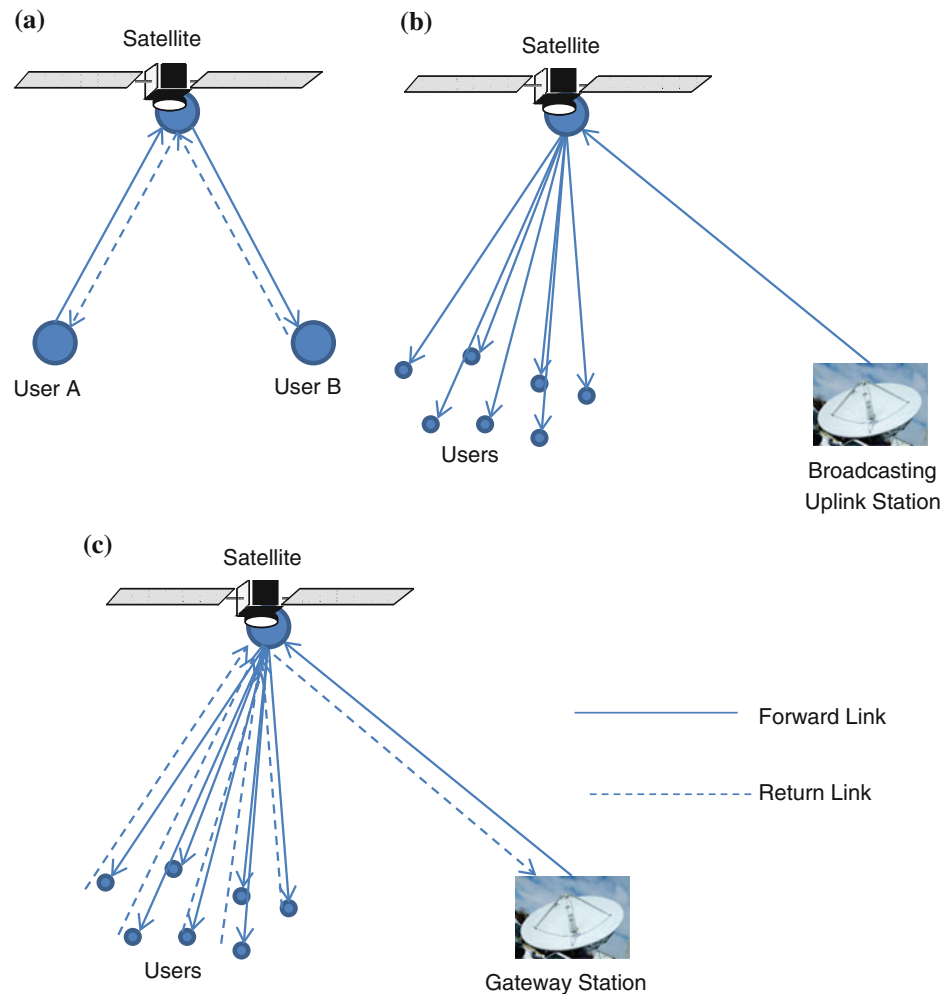
downlink beams [7]. Such architectures are typically used in mobile satellite systems.

The main advantage of regenerative processing payloads over bent-pipe payloads is the fact that processing payloads separate the uplink and downlink noise signals, thus improving the overall satellite link quality. In addition, the ability to route and switch signals on-board the satellite provides a significant advantage for satellite networks such as Spaceway. These advantages are at the expense of a much more complex payload, heavier mass, and greater power consumption. In addition, regenerative processing payloads are uniquely designed to operate with specific communications protocols, thus limiting their operational capabilities to these specific protocols.

Another architectural aspect for communications satellites is how users are connected to other users, or to information sources. Figure 14.4 illustrates the three common connectivity architectures for communications satellites

1. Two way symmetric or 'one-to-one' (Fig. 14.4a), where the amounts of information exchanged between any pair of users is almost equal. Each user terminal has both receiving and transmitting capability. The network interconnections can be a mesh topology (i.e. the nodes/terminals in the same network can communicate with each other via a single relay link through the satellite) or a star topology (i.e. all signal transmissions to and from an individual node/terminal must be routed through a central

Fig. 14.4 Three common connectivity architectures for communications satellites. **a** Two way architecture symmetric one to one VSATs. **b** Broadcasting architecture one way (one to many). **c** Multimedia internet access architecture two way unbalanced (many to one)



location or hub via multiple relay links through the satellite). The mesh network connectivity can be established either by ground hub control stations for bent-pipe satellites, or by on-board processors for processing payloads.

2. Broadcasting or 'one-to-many' (Fig. 14.4b), where information flows one way from the broadcasting transmitting station to all of the receiving users. The users terminals have only receive capability.
3. Multimedia Internet Access Architecture (Fig. 14.4c), where two-way unbalanced or 'many to one' interconnectivity is used. The users terminals have both receive and transmit capability. The forward links (gateways to users) have much higher capacity/bandwidth than the return links (users to gateways).

14.3 Communications Link and Performance

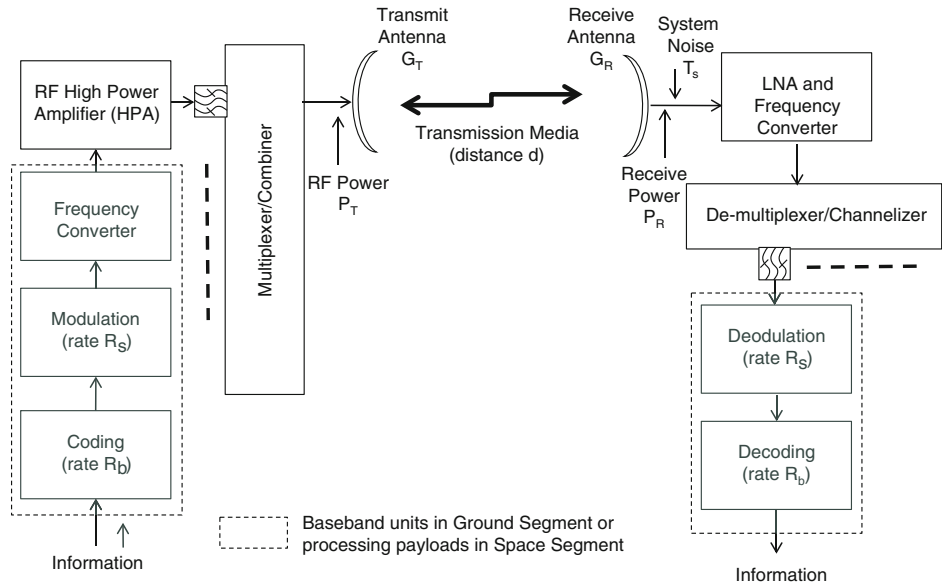
A satellite communications system consisting of a space segment and a ground segment is depicted in Fig. 14.1. For any satellite mission to be functional, control and/or mission

information has to be exchanged between the two segments through the communications link in the form of information bearing radio frequency (RF) signals. As discussed in the following sections, the performance of a communications link is primarily determined by (a) the RF signal power transmitted in the direction of the receiver, (b) the loss of RF energy in the transmission media, and (c) the ability of the receiver to convert the incoming RF signal from a given direction into electrical energy.

The control information (i.e. the telecommands) usually originates from the TT&C station. The mission information can be originated either from the ground segment, as it is for the communications satellite systems, or from the space segment, as it is for the sensor/exploring satellite systems.

A complete satellite communications link is composed of one or more link segments, such as an uplink from a ground station to a satellite, a downlink from a satellite to a ground station, and/or an inter-satellite link. The radio frequency bands commonly used for space communications links are summarized in Table 14.1.

Fig. 14.5 A typical communications link



14.3.1 Communications Link Elements

Generally speaking, there are three major elements in any communications link, including satellite uplink and downlink, as shown in Fig. 14.5: a transmit system, the transmission media, and a receive system.

14.3.1.1 Transmit System

The transmit system for a satellite uplink is an Earth station and for a satellite downlink it is the satellite. The primary function of a transmit system is to amplify an information bearing RF signal to a proper level through a high power amplifier (HPA), and then to radiate the signal to transmission media (atmosphere and space) through a transmit antenna. The transmit system in an Earth station or a processing satellite usually also includes the units for encoding the baseband source information and modulating it onto an information bearing RF signal prior to routing that to the HPA for high power amplification.

An important figure of merit of a transmit system is its equivalent isotropically radiated power (EIRP) which is a product of the transmit antenna gain (G_T) and the RF signal power delivered by the high power amplifier to the antenna (P)

$$EIRP = G_T \cdot P_T. \quad (14.1)$$

The EIRP is usually expressed as $10 \log(P) + 10 \log(G_T)$ and then has units of dBW; i.e. dB relative to 1 W. For a given set of source encoding and modulation parameters, including forward error correction in order to provide signal redundancy, the EIRP directly influences the link performance and the capacity of the communications system. It is one of the most important parameters in the

design considerations and trade-offs for any space communications system due to its critical influence on the communication link performances/capability and to the expensive nature of the DC power in a spacecraft. The details of microwave antennas and high power amplifiers are described later in this chapter.

14.3.1.2 Receive System

The receive system for the uplink is the satellite and for the downlink it is the Earth terminal. The receive system accepts the information-bearing RF signal through a receive antenna and passes it to a low noise amplifier to minimize the noise contributions of downstream circuitries. When the receive system is the Earth terminal, it demodulates the RF signal into the baseband form suitable for end users. A figure of merit for a receive system, whether it be the satellite or the Earth terminal, is the ratio of its antenna gain (G_R), expressed as a numerical ratio, to its system noise temperature (T_S) in kelvins. The gain to noise temperature ratio (G/T) is usually expressed as the dB difference of antenna gain in dBi (dB with respect to isotropic) and noise temperature in dBK, as $G/T(\text{dB/K}) = 10 \log(G_R) - 10 \log(T_S)$. For a given set of transmission parameters, the G/T governs the signal-to-noise ratio (S/N) at the input of the receive system and, consequently, significantly influences the performance and capacity of the overall link.

The system noise temperature, measured in kelvins, is due to an effective noise temperature (T_e) generated by the internal sources such as the receiver (i.e. the low noise amplifier and frequency converter) and the input passive components, as well as by an antenna noise temperature (T_a) generated from external sources in the field of view of

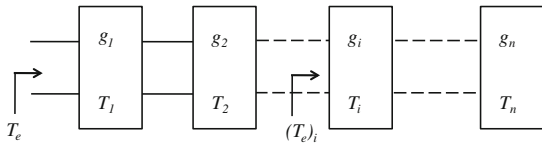


Fig. 14.6 Noise in a cascaded network

the receive antenna. For a given signal bandwidth B in Hz, the power due to the thermal noise is given by

$$N = N_o B = kT_s B = k(T_a + T_e)B \quad (14.2)$$

in watts, where $N_o = kT_s$ is the noise power density measured in W/Hz and $k = 1.3806505 \times 10^{-23}$ J/K is the Boltzmann constant. In practical applications, the thermal noise power is the ultimate limit to the communications performance and capacity for any communications system based on Shannon's information theorem [8].

14.3.1.3 Noise from Internal Sources

A receive system is composed of many active and passive components (including connecting lines). The equivalent noise temperature is defined at the input of each component. For an active component

$$T_r = T_0(F - 1) \quad (14.3)$$

where F is the *noise figure* of the active component and T_0 is a normal ambient temperature of 290 K based on the *noise figure* definition. And for a passive component

$$T_L = T(L - 1) \quad (14.4)$$

where L is the loss of the passive component expressed as the numerical ratio of the input power to the output power ($L \geq 1$) and T is the physical temperature of the passive component. For a general cascaded network with n active/passive components, each has gain of g_i (for a passive component, $g = 1/L$) and the equivalent noise temperature of T_i as shown in Fig. 14.6, the total noise power that is generated by all the components and which appears at the output of component n is

$$N_n^{out} = kT_1 B \prod_{i=1}^n g_i + kT_2 B \prod_{i=2}^n g_i + \cdots + kT_n B g_n. \quad (14.5)$$

The equivalent noise power of the overall cascaded network referenced at the input of component 1 can be obtained as

$$N_1^{in} = kT_e B = \frac{N_n^{out}}{\prod_{i=1}^n g_i} = kB \left(T_1 + \frac{T_2}{g_1} + \cdots + \frac{T_n}{\prod_{i=1}^{n-1} g_i} \right). \quad (14.6)$$

Therefore, the overall equivalent noise temperature (T_e) from all the internal sources of a cascaded network, referenced at the input of the very first component, is

$$T_e = T_1 + \sum_{i=2}^n \left(T_i / \prod_{j=1}^{i-1} g_j \right). \quad (14.7)$$

In fact, the equivalent noise temperature may be referenced at any point in the cascaded network. For example, the equivalent noise temperature of the network referenced at the input of component i can be derived as

$$(T_e)_i = \sum_{j=1}^{i-1} \left(T_j \times \prod_{k=j}^{i-1} g_k \right) + T_i + \sum_{j=i+1}^n \left(T_j / \prod_{k=i}^{j-1} g_k \right). \quad (14.8)$$

With this concept, it can be easily proved that the G/T of a receive system is independent of the reference point of the G/T calculation.

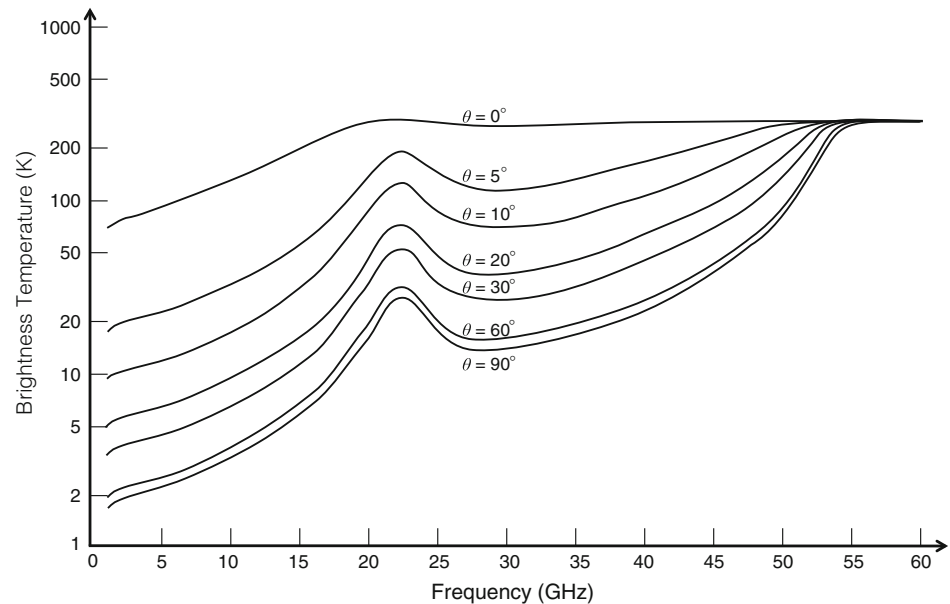
14.3.1.4 Noise from External Sources

The noise collected from the noise sources in the field of view of the receive antenna is the dominant contributor to the antenna noise temperature. The antenna noise temperature can be expressed as the convolution of the antenna pattern $g(\Omega)$ and the brightness temperature $T(\Omega)$ of external noise sources as follows [9]

$$T_a = \frac{1}{4\pi} \iint g(\Omega) \cdot T(\Omega) \cdot d\Omega \quad (14.9)$$

where $d\Omega$ represents an infinitesimal element of solid angle in the direction of the noise source in the antenna coordinate system. The external noise sources include the stars, background cosmic noise, and absorbent media such as atmospheric attenuation. For a communications link between the ground segment and the space segment, the antenna noise temperature is dominated by the Earth brightness temperature and/or the atmospheric brightness temperature depending on the main direction of the receive antenna beam. For a downlink, the receive antenna (an Earth station antenna) is pointed 'upward' to a spacecraft and the main noise source is the atmospheric attenuation (i.e. the sky brightness temperature). For an uplink, the receive antenna (a satellite antenna) is pointed toward Earth and the main noise source is the Earth's emissivity (i.e. the Earth's brightness temperature). Figure 14.7 shows an example of the sky brightness temperature for an Earth station antenna at different elevation angles [10] in a 'clear sky' condition. As can be seen, an Earth station antenna noise temperature is largely dependent on the elevation angle and its operating frequency. For an Earth station antenna with a narrow beamwidth and a higher elevation angle, the antenna pattern will be largely encompassed by the atmosphere and therefore the sky brightness temperature will be the same as the antenna noise temperature.

Fig. 14.7 Brightness temperature for clear air for 7.5 g/m³ of water vapor concentration for different Earth station antenna elevation angles θ [10]



As will be discussed later, rain attenuation can be significant at high frequency. A byproduct of rain attenuation is an increase of the emission noise, and therefore an increase of the Earth station antenna noise temperature. The additional antenna noise temperature increase from the ‘clear sky’ case as compared to the ‘output’ of the attenuating medium at the antenna aperture can be estimated by

$$\Delta T_a = T_m(1 - 10^{-A/10}) \tag{14.10}$$

where A is the rain attenuation in dB and T_m is the effective temperature of the medium, which lies in the range 260–280 K at frequencies between 10 and 30 GHz.

The noise temperature of an Earth-pointing spacecraft antenna is determined by the location of the spacecraft, the antenna pointing direction, the antenna coverage area on Earth, and the operating frequency. As an illustration, Fig. 14.8 provides the antenna noise temperature of an Earth coverage antenna on a geostationary spacecraft at different orbital slots and operating frequencies. The variation of the antenna noise temperature over orbital slots for a given frequency is due to the relative ratios of land mass and ocean in the field of view of the antenna.

Under certain operational conditions, part of the main beam of the receive antenna could see one or more very bright noise sources such as the Sun (with a brightness temperature around 10,000 K). In this case, the antenna noise temperature can increase significantly and cause temporary communications service outages. These conditions occur at the equinoxes when the Sun transits the position of a geostationary satellite as seen from the Earth.

For an inter-satellite link, the receive antenna is pointed to another satellite and the majority of its antenna beam is directed towards ‘cold’ space and will see the background

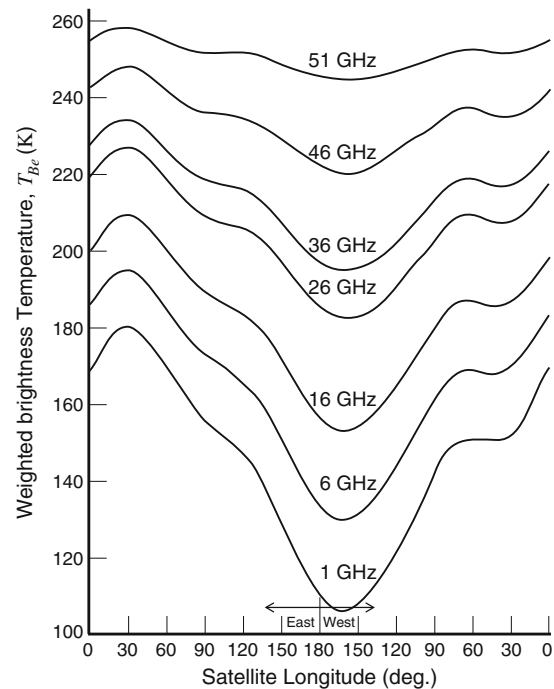


Fig. 14.8 Earth coverage antenna noise temperature as a function of longitude viewed from geostationary orbit at frequencies between 1 and 51 GHz [10]

cosmic noise at around 2.7 K. This usually results in a very low antenna noise temperature.

14.3.1.5 Transmission Media

Unlike an inter-satellite link in which the radio signal is transmitted in free space, the radio wave in an uplink or a downlink propagates through the atmosphere and

Table 14.2 Estimated ionospheric effects for elevation angles of about 30° one-way traversal [9]

Effect	Frequency dependence	0.1 GHz	0.25 GHz	0.5 GHz	1 GHz	3 GHz	10 GHz
Faraday rotation	$1/f^2$	30 rotations	4.8 rotations	1.2 rotations	108°	12°	1.1°
Propagation delay	$1/f^2$	25 μ s	4 μ s	1 μ s	0.25 μ s	0.028 μ s	0.0025 μ s
Refraction	$1/f^2$	<1°	<0.16°	<2.4'	<0.6'	<4.2''	<0.36''
Variation in the direction of arrival (r.m.s)	$1/f^2$	20'	3.2'	48''	12''	1.32''	0.12''
Absorption (auroral and/or polar cap)	$\approx 1/f^2$	5 dB	0.8 dB	0.2 dB	0.05 dB	6×10^{-3} dB	5×10^{-4} dB
Absorption (mid-latitude)	$1/f^2$	<1 dB	<0.16 dB	<0.04 dB	<0.01 dB	<0.001 dB	< 1×10^{-4} dB
Dispersion	$1/f^3$	0.4 ps/Hz	0.026 ps/Hz	0.0032 ps/Hz	0.0004 ps/Hz	1.5×10^{-5} ps/Hz	4×10^{-7} ps/Hz
Scintillation	See Rec. ITU-R P.531	See Rec. ITU-R P.531	See Rec. ITU-R P.531	See Rec. ITU-R P.531	>20 dB peak-to-peak	≈ 10 dB peak-to-peak	≈ 4 dB peak-to-peak

Scintillation values observed near the geomagnetic equator during the early night-time hours (local time) at equinox under conditions of high sunspot number

ionosphere, and therefore is subjected to propagation impairments.

Caused by solar radiation, the Earth's ionosphere consists of several regions of ionization in the upper atmosphere. The total electron content (TEC) accumulated along the transmission path penetrating the ionosphere causes a rotation of the polarization (i.e. Faraday rotation) of the propagating radio wave, a time delay of the signal, a change in the apparent direction of arrival due to refraction, and scintillations. The ionospheric effects for a high value of TEC and an elevation angle of 30° one-way traversal is given in Table 14.2. The ionosphere effects above 10 GHz are negligible; however, the effects can be significant for non-geostationary satellite services below 3 GHz.

The main effects of the non-ionized atmosphere are due to signal absorption by atmospheric gases (including humidity) and rainfall. Attenuation by atmospheric gases is mainly due to oxygen and water vapor absorptions and is dependent mainly on frequency, elevation angle, altitude above sea level, and water vapor density (absolute humidity). A detailed atmospheric attenuation model can be found in [11]. As an illustration, Fig. 14.9 presents the attenuation by atmospheric gases at an elevation angle of 90° at sea level. As shown in the figure, the atmospheric attenuation is generally insignificant at frequencies below 10 GHz. Above 10 GHz, there exist several 'windows' with relatively small attenuation and several attenuation peaks. The attenuation peaks of dry air and water vapor are due to interactions of the electromagnetic field of the wave with the magnetic moment of oxygen molecules and the polar (electric dipole) molecules of water vapor, respectively.

The frequency spectrums in the 'windows' are selected for the uplinks and downlinks of space-Earth

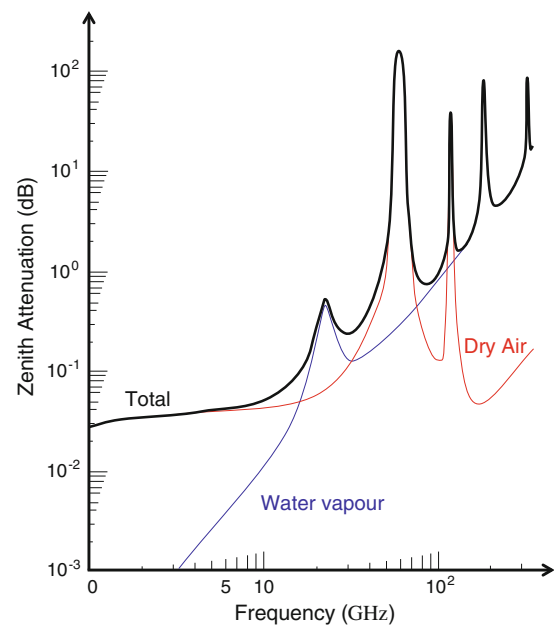
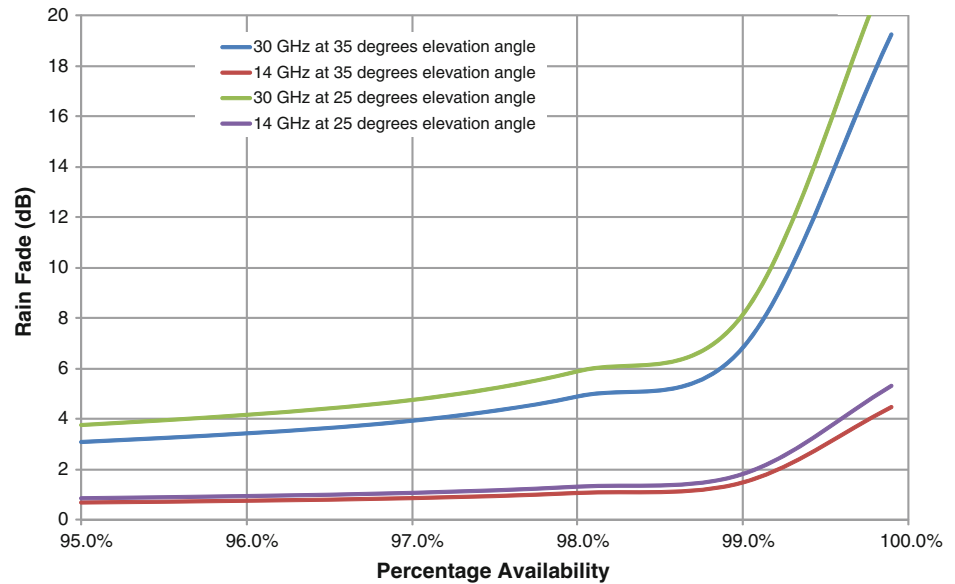


Fig. 14.9 Total, dry-air and water-vapor zenith attenuation from sea level with surface pressure of 1,013 hPa, surface temperature of 15 °C, and surface vapor of 7.5 g/m³ [11]

communications. The frequency spectrums at the attenuation peaks, such as the 60 GHz band, are appropriate selections for inter-satellite links because the high atmospheric attenuation can be used as an advantage to shield the link from ground originated interferences and/or jams.

The absorption due to rain becomes significant at frequencies exceeding 10 GHz, when the wavelength approaches the size of raindrops, and therefore is an important consideration when designing a space-Earth link for a

Fig. 14.10 Rain attenuation predicted by ITU-R model



higher frequency. Rain attenuation is also strongly dependent on elevation angle (path length in rain) and rainfall rate distribution. The rainfall rate distribution varies with time and geographic locations. In the absence of available global-wide experimental data, many global rain attenuation prediction models have been developed based on the available test results. The most commonly used models are the Crane model [12, 13] and the ITU-R model [14]. As an example, Fig. 14.10 shows the rain attenuation versus frequency at different elevation angles for rainfall rates that exceed 5–0.1 % of the time. The figure, generated for a location in North America based on the ITU-R model, clearly shows that the rain attenuation is strongly dependent on the frequency, the elevation angle, and the time percentage for the rain rate (which is directly associated to link availability as shown in a later section). A byproduct of rain attenuation is an increase of antenna noise temperature, as indicated in Eq. 14.10.

14.3.2 Modulation, Coding and Multiple Access

14.3.2.1 Modulation

A communications link can be used to transmit voice, video, and/or data information. The information such as voice and video data are baseband signals that are encoded, in some cases multiplexed with other signals, and modulated to superimpose them upon a high frequency carrier signal that can be radiated efficiently by the antenna.

Modulation can take several forms including varying the signal envelope, and/or phase (or frequency) in the RF for efficient transmission. Dependent on the signal type,

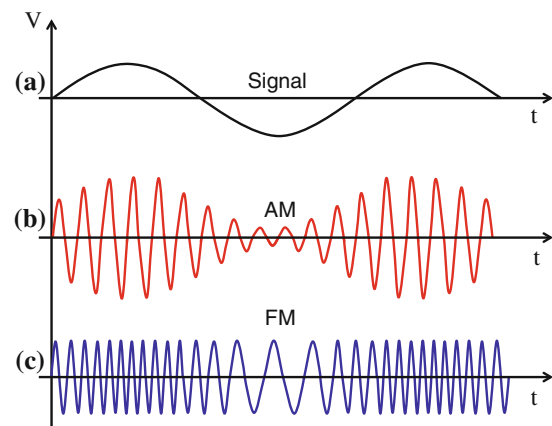
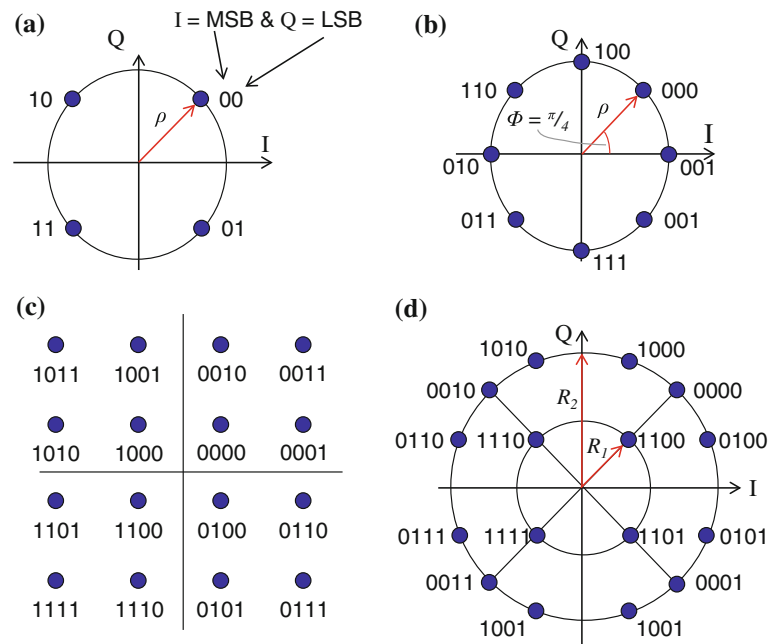


Fig. 14.11 Analog modulation **a** a modulating baseband signal; **b** an amplitude modulated signal; **c** a frequency modulated signal

modulation can be divided into analog modulation and digital modulation [15, 16]. In an analog modulation, the baseband signal may be superimposed on a carrier signal to vary its amplitude, frequency, or phase for amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM), respectively, as illustrated in Fig. 14.11.

In digital modulation, the baseband data with a data rate of R_b bits/s can be mapped into M states in amplitude and/or phase/frequency of the RF carrier signals to form an M -ary modulation, where $M = 2^n$ and n is the number of bits represented by each state of the modulated carrier. The group of bits transmitted in each of the M states is a symbol. The most common methods of digital modulation for satellite communications are M -ary frequency shift keying (MFSK), M -ary phase shift keying (MPSK), and M -ary quadrature amplitude modulation (MQAM). Recently, a super-class of

Fig. 14.12 Constellation of
a QPSK; **b** 8PSK; **c** 16QAM; and
d 16APSK



MQAM modulation called asymmetric phase-shift keying (APSK) was introduced in the DVB-S2 standard [6] for improved performance in a nonlinear satellite transponder.

An MFSK modulator puts the frequency of the carrier into one of M frequencies (separated by $1/(2T_S)$, where T_S is the duration of a symbol) according to the value of a modulated voltage. The M transmitted signals are of equal energy, of equal duration, and are orthogonal to each other. The bandwidth efficiency of a coherent MFSK is [17, 18] $B_E = R_b/B = (\log_2 M)/M$, which decreases with increasing M . Therefore, MFSK modulation is bandwidth inefficient. However, an MFSK signal is power efficient as it is a constant envelope modulation (CEM) and its signal is insensitive to nonlinearity in the high power amplifier in the transmitter (i.e. the amplifier can be operated at saturation). In addition, a CEM signal is more tolerant of random noise as well as Rayleigh fading. Due to this feature, FSK has been used to modulate the tele-command/control signals for many satellites and spacecraft.

An MPSK modulator puts the phase of the carrier into one of M states according to the value of a modulated voltage. The modulated waveform expressed as two states or bi-phase PSK is called BPSK, and four states or quadriphase is termed QPSK. In MPSK modulation, the symbol states are equally spaced on a cycle of constant symbol energy. As an illustration, Fig. 14.12a and b show the constellations of QPSK and 8PSK, respectively. Obviously, an MPSK without pulse shaping is a constant envelope modulation. However, in RF communication, pulse shaping is essential for making the signal fit in its frequency band. The ratio of the data rate to the minimum bandwidth assuming ideal Nyquist filtering is $B_E = R_b/B = \log_2 M$.

Unlike MFSK, the bandwidth efficiency of an MPSK signal increases with increasing M . However, with increasing M , the constellation is more crowded, resulting in reduced tolerance of noise tolerance and therefore decreased power efficiency.

An MQAM modulator puts the carrier vector formed by different phase and amplitude into M states according to the value of a modulated voltage, which can be considered as an MPSK with variable amplitudes. Figure 14.12c shows the constellation diagram of a 16QAM. The bandwidth efficiency of an MQAM signal is identical to MPSK. Unlike MPSK, the MQAM symbols are unequally spaced and do not have constant symbol energy. When M is large, the symbol spacing of MQAM is less crowded than MPSK, and therefore an MQAM signal is less sensitive to noise and interference than an MPSK modulated signal.

APSK can be considered as a super-class of quadrature amplitude modulation. The constellation of a 16APSK is shown in Fig. 14.12d. The advantage over conventional QAM, for example 16QAM, is a lower number of possible amplitude levels, resulting in fewer problems with nonlinear amplifiers.

The selection of the modulation scheme for satellite communications is done according to power and bandwidth efficiency. The power efficiency is the ratio of the required bit energy (E_b) to noise spectral density (N_o) for a certain bit error probability (P_b) of digital communication over an additive white Gaussian noise (AWGN) channel for a given modulation scheme. Bandwidth efficiency is the ability to accommodate data within a limited bandwidth of a channel.

The trade-off between bandwidth efficiency and power efficiency of different modulation schemes is depicted in

Fig. 14.13 Bandwidth and power efficiency of different digital modulations [39, 40]

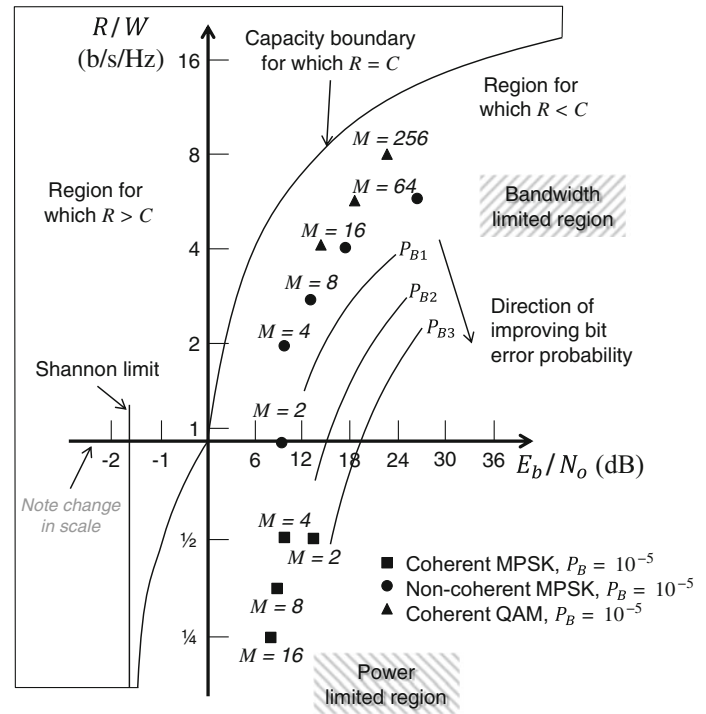


Fig. 14.13, from which it is clear that BPSK and QPSK have the same power efficiency, but QPSK has twice the bandwidth efficiency. MFSK has better power efficiency but poorer bandwidth efficiency. MPSK and MQAM have the same bandwidth efficiencies but MQAM has better power efficiency when M is greater than eight.

BPSK and QPSK were the most commonly used modulation schemes for satellite communications for balanced power and bandwidth efficiencies, but with the ongoing demand of efficient use of the spectrum for high speed satellite communications, including the high definition DTH television services, 8PSK, 8QAM, 16QAM, and 16APSK are now becoming common in commercial satellite links, benefiting from the significant progress in the coding and error correction theory.

14.3.2.2 Coding and Error Correction

Noise, interference, and distortion of a digital communications link will occasionally cause bits to be missed or misinterpreted by the receiving system and therefore increase the bit error rate (BER) which is defined as the ratio of error bits to the total bits in a given time interval. For given information rate and the transmission bandwidth, a natural way to reduce BER is to increase the bit energy (E_b) relative to the noise/interference (N_o); i.e. to increase E_b/N_o by increasing the transmit EIRP. However, this approach is not always practical, particularly for satellite communications where such factors as cost and interference

place limits on the EIRP. The preferred way is to reduce the threshold of E_b/N_o while keeping the bandwidth efficiency as high as possible. This goal may be achieved by error correction. Error correction may generally be realized in two different ways:

- Automatic repeat request (ARQ) (also referred to as backward error correction): This is an error control technique whereby an error detection scheme is combined with requests for retransmission of erroneous data.
- Forward error correction (FEC): The transmitter system encodes the data using an error-correcting code (ECC) prior to transmission. In the encoding process, redundant bits are added to an incoming bit stream so that errors in transmission may be detected and corrected at the receiver using a sophisticated decoder. FEC uses no information feedback to the transmit side, therefore there is no additional delay introduced by the error correction process. FEC is commonly used in satellite communications to overcome the effects of interference in a satellite link.

In an FEC process, the encoder accepts binary information at a rate of R bits/s and generates encoded binary data at a rate of R_b with the coding rate of r ($r < 1$), where $R_b = R/r$, by introducing redundant bits for error detection and correction.

The two main classes of FEC codes are convolutional codes and block codes. Convolutional codes work on bit or symbol streams of arbitrary length. The encoded data depends not only on the most recent data but also on a

specified number of previous source data, so that a sliding sequence of past data bits is used to generate several encoded data bits. Viterbi decoding is mostly used for convolutional codes. Block codes work on fixed-size blocks (packets) of bits or symbols of predetermined size. Reed-Solomon (R/S), codes, BCH codes (developed in 1959 by Alexis Hocquenghem (c. 1908–1990), and independently in 1960 by Raj Chandra Bose (1901–1987) and Dwijendra Kumar Ray-Chaudhuri (born 1933), the acronym BCH comprises the initials of these inventors' names), Hamming codes, and Golay codes, are a few examples of commonly used block codes. The highly efficient low-density parity-check (LDPC) codes belong to linear block codes.

Classical block codes and convolutional codes may be combined to form more powerful codes to provide near Shannon-limit performance. For example, the Viterbi codes and the Reed-Solomon codes are combined in concatenated coding schemes in DVB-S, and LDPC and BCH are combined to form new codes in DVB-S2 [6] for high-speed satellite communications applications. Table 14.3 compares the bandwidth and power efficiencies of several modulation and coding schemes. As can be seen, in comparison to an uncoded QPSK signal, the power efficiency of the signal with QPSK modulation and $\frac{3}{4}$ -Viterbi coding scheme improves the power efficiency by about 2.5 dB (60 %) while the bandwidth efficiency is reduced by 25 %. With the LDPC/BCH codes in DVB-S2, the power efficiency can be further improved by 4.8 dB without significantly decreasing the bandwidth efficiency. In fact, the powerful coding schemes, such as the LDPC/BCH codes and Turbo codes, have made it possible for digital satellite communications to use more bandwidth-efficient modulations with the throughput for a given bandwidth close to the Shannon limit as it is depicted in Fig. 14.14.

To further improve the bandwidth usage particularly in the high fading link, adaptive coding and modulation (ACM) techniques have been proposed and adapted in wideband satellite communications and MSS satellite communications in order to provide very significant increases in capacity [6]. ACM involves managing the modulation level and coding rate for each user terminal based on the instantaneous link fading condition. As the fading condition changes for each individual terminal, the modulation level and code rate are changed in order to maintain the BER requirements. As only a low percentage of user terminals in a service area will encounter large fading at any time, this technique significantly increases the average information throughput per unit bandwidth.

These E_b/N_o performances are based on an ideal communications channel. In reality, the nonlinearity, group delay variation, gain flatness, phase noise, and interference existing in the communications channel will greatly impact the communications link performances.

Table 14.3 A comparison of bandwidth and power efficiencies of several modulation and coding schemes

Modulation	Coding	E_b/N_o (dB) BER < 10^{-5}	Bandwidth efficiency
QPSK	None	9.6	2.0
QPSK	Viterbi, $r = \frac{3}{4}$	7.1	1.5
QPSK	DVB-S Viterbi $r = \frac{3}{4} + \text{R/S}$	5.0	1.38
QPSK	DVB-S2 LDPC + BCH; $r = \frac{3}{4}$	2.3	1.49
8PSK	DVB-S2 LDPC + BCH; $r = \frac{3}{4}$	4.4	2.23
16APSK	DVB-S2 LDPC + BCH; $r = \frac{3}{4}$	5.4	2.97

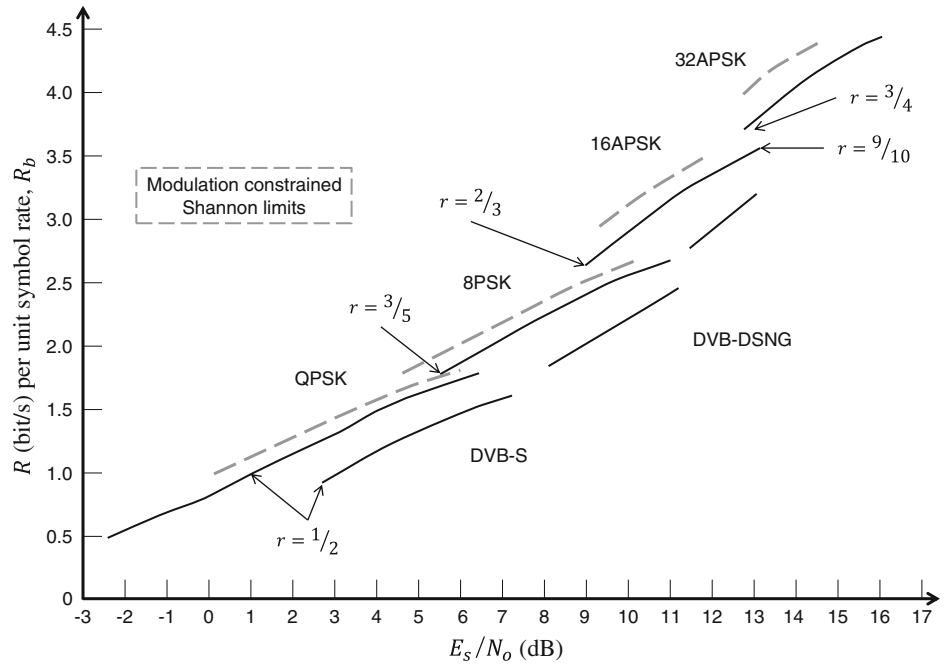
14.3.2.3 Multiplexing and Multiple Access

Almost all communications satellites require sharing of the satellite link resource/capacity among multiple users. This ability, allowing access to a given user by a community of users through combining the respective individual signals into a single one, is accomplished by multiplexing techniques, while the ability for multiple users to transmit their respective data through a given satellite transponder is realized by multiple access techniques. The commonly used multiplexing techniques in satellite communications include

- Frequency division multiplexing (FDM): This combines multiple signals with non-overlapping frequency bands into one wideband signal for transmission. Each signal can be recovered by filtering at the corresponding user receivers.
- Time division multiplexing (TDM): Each signal is compressed into a high speed signal and the multiple signals are transmitted at the same frequency but in different time slots. The signal is recovered at the respective receiver by selection of the specific time slot in which its signal was transmitted.
- Code division multiplexing (CDM): Each signal is assigned a unique signature code chosen from a set of orthogonal codes before they are combined in the frequency and time domain. The signal can be recovered by cross-correlating the signal with the identical signature code generated by the respective receiver.

Similarly, the commonly used multiple access methods in satellite communications are frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA). In many applications, a combination of more than one multiplexing techniques and multiple access techniques is adopted for most efficient use of the satellite resources. A summary of various multiple access methods and multiplexing methods in satellite communications is given in Table 14.4.

Fig. 14.14 Required E_s/N_o versus spectrum efficiency in the additive white gaussian noise (AWGN) channel with DVB-S2 modulation/coding schemes (ideal demodulator), $E_b/N_o = E_s/N_o - 10\log(B_e)$



14.3.3 Communications Link Design and Performance Analysis

The fundamental objective of a satellite communications link design is to determine the technical parameters of the satellite transmit system (i.e. EIRP) and receive system gain to noise temperature ratio (G/T) in order to maximize the information to be transmitted in a given bandwidth to meet the system availability and quality requirements.

For a typical communications link as, shown in Fig. 14.15, with the transmit EIRP defined as per Eq. 14.1, the power flux density (PFD) at the receive antenna, which is d meters away from the transmit antenna, is

$$PFD = \frac{EIRP}{4\pi d^2} \cdot \frac{1}{A} = \frac{G_T P_T}{4\pi d^2} \cdot \frac{1}{A} \quad (14.11)$$

in W/m^2 , where $4\pi d^2$ is often referred to as the spreading loss and A is the attenuation induced from the transmission medium including the atmospheric attenuation, cloud/rain attenuation, and other propagation attenuations. Note that the PFD is independent of frequency for a fixed transmit antenna gain.

The received signal level at the output of the receive antenna can be obtained as

$$\begin{aligned} P_R &= PFD \cdot G_R \frac{\lambda^2}{4\pi} = P_T G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2 \cdot \frac{1}{A} \\ &= P_T G_T G_R \cdot \frac{1}{\rho_L} \cdot \frac{1}{A} \end{aligned} \quad (14.12)$$

in W , where λ is the wavelength of the carrier signal in meters and $\rho_L = (4\pi d/\lambda)^2$ represents the power loss

between two isotropic antennas that is often referred as the frequency-dependent path loss.

For a receive system with a total system noise temperature of T_s and a receive bandwidth of B Hz, the received carrier to noise ratio can be obtained as

$$\frac{C}{N} = \frac{P_R}{N} = P_T G_T \cdot \frac{G_R}{T_s} \cdot \frac{1}{\rho_L} \cdot \frac{1}{A} \cdot \frac{1}{kB} \quad (14.13)$$

In link analysis, this equation is usually expressed in decibels as

$$\begin{aligned} \frac{C}{N} (dB) &= P_T (dBW) + G_T (dBi) + G_R (dBi) - 10 \log T_s \\ &\quad - \rho_L (dB) - 10 \log A - 10 \log B + 228.6. \end{aligned} \quad (14.14)$$

In addition to the thermal noise, interference contributions must be taken into account in link analysis for a satellite communication system. These include

- Interference generated by the same satellite due to frequency reuse: for frequency reuse through the orthogonal polarization, the interference is due to the finite cross-polarization isolation of the satellite antenna. The typical carrier-to-interference ratio (C/I) due to cross-polarized co-frequency interference is about 27 dB. For frequency reuse through spatial diversity, such as the multi-spot beams used in S-band mobile satellite services and Ka-band satellite services, the C/I is more complicated, being heavily dependent on the beam-frequency mapping schemes and the beam shaping. A typical aggregated C/I for this case is between 14 and 20 dB.

Table 14.4 A summary of multiple access and multiplexing in satellite communications

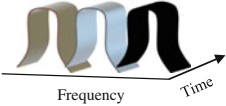
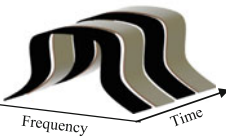
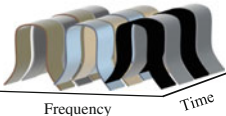
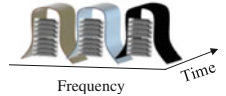
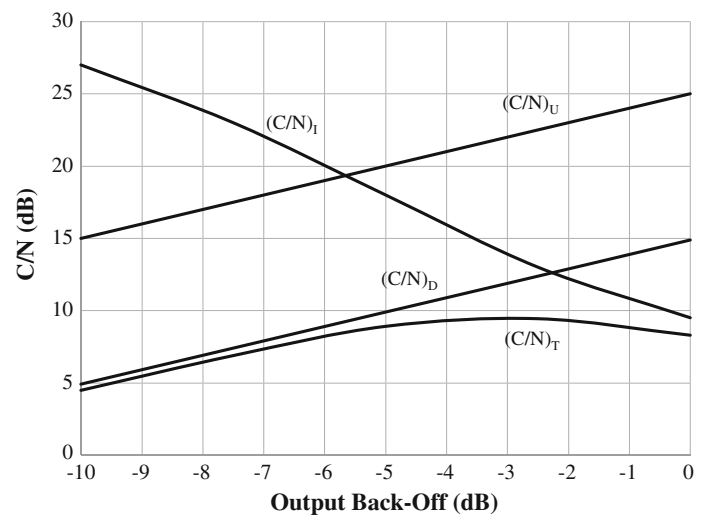
Multiple access	Multiplexing	Advantage	Disadvantage
FDMA 	FDM	<ul style="list-style-type: none"> • No time reference • Simple implementation • Low or medium traffic 	<ul style="list-style-type: none"> • Satellite transponder back-off required for linear operation • User terminal/Modem requires multiple filtering • Poor flexibility for traffic reconfiguration • May require uplink power control
TDMA 	TDM	<ul style="list-style-type: none"> • Small or no back-off for satellite transponder • High traffic for large number of users • Easy traffic reconfiguration and demand assignment 	<ul style="list-style-type: none"> • Require network synchronization • High peak EIRP requirement for all E/S • Require memory buffers
FDMA + TDMA (shared transponder) 	FDM + TDM	<ul style="list-style-type: none"> • Easy traffic reconfiguration and demand assignment • Low or medium traffic 	<ul style="list-style-type: none"> • Satellite transponder back-off • Require network synchronization for TDMA E/S need • High peak EIRP requirement for all E/S • Require memory buffers for TDMA E/S
FDMA + CDMA 	FDM + CDM	<ul style="list-style-type: none"> • Simple implementation • Low EIRP requirement for E/S and/or user terminal • Less sensitive to multi-path interference 	<ul style="list-style-type: none"> • Limited traffic capacity due to Multiple Access Interference • Satellite transponder back-off • Require synchronization

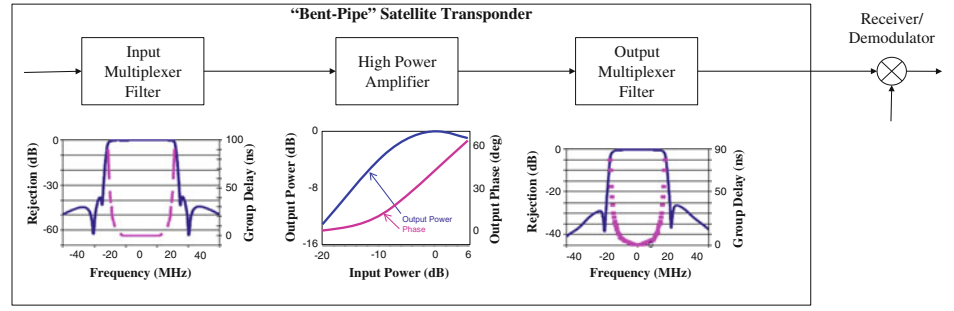
Fig. 14.15 A typical link budget optimization for FDMA signal



- Interference due to the inter-modulation products or the noise power ratio (NPR, defined as the ratio of the total output power to the uncorrelated in-band distortion power) caused by the nonlinearities in transmit systems in a multi-carrier operation condition. To minimize the system impact, the inter-modulation/NPR is usually traded off with the output back-off of the high power

amplifier. For a satellite transmit system in a multi-carrier operation, the back-off is usually around 3 dB which yields about 15 dB NPR. For a ground transmit system, the high power amplifier back-off is usually 5–6 dB providing better than 20 dB C/I. The trade-off between the back-off and the inter-modulation interference for a typical high power amplifier is illustrated later in Fig. 14.54.

Fig. 14.16 A simplified satellite transponder and typical characteristics of the major components



- Emissions received from an adjacent satellite by sidelobes of Earth stations of the concerned satellite (downlink), and emissions received from an adjacent satellite uplink Earth station by the concerned satellite receive antenna sidelobes (uplink) [9].

The common method for dealing with various interferences is to include it all as additional noise so that the overall link C/N can be obtained, in a numerical value, as

$$\begin{aligned} \left(\frac{C}{N}\right)_T &= \left[\left(\frac{C}{N}\right)^{-1} + \left(\frac{C}{I}\right)^{-1} \right]^{-1} \\ &= \left[\left(\frac{C}{N}\right)^{-1} + \sum_i \left(\frac{C}{I_i}\right)^{-1} \right]^{-1} \end{aligned} \quad (14.15)$$

where C/I_i is the itemized carrier to interference ratio and C/I is the aggregated carrier to interference ratio.

For a digital communications link with an information rate of R bits/s and a symbol rate of R_s symbols/s, the bit energy to noise density ratio can be obtained from the carrier to noise ratio using the relationships of $C = E_b R$ and $B = R_s$ as

$$\frac{E_b}{N_0} (\text{dB}) = \left(\frac{C}{N}\right)_T (\text{dB}) - 10 \log \left(\frac{R_s}{R}\right). \quad (14.16)$$

For a spacecraft with a regenerative on-board system that demodulates the uplink signal to baseband and then reroutes it, the uplink and downlink noise contributions are somewhat separated. In that case, the overall bit error rate $(BER)_T$ of the system can be approximated from the uplink BER and the downlink BER [9] as

$$(BER)_T = (BER)_U + (BER)_D. \quad (14.17)$$

To meet the overall system BER requirement, the BER for the uplink and the downlink have to be individually minimized. For the case where the E_b/N_0 performances between the uplink and the downlink are equal, the BER of each link should be one-half of the system BER. However, with the advanced encoding/modulation technique used in modern digital satellite communications, the reduced BER will have no significant impact on the required E_b/N_0 for the uplink and the downlink respectively, and they can be considered to be independent communication links.

For a bent-pipe communications satellite, the overall link $(C/N)_T$ is a combination of the uplink $(C/N)_U$, the downlink $(C/N)_D$, and all the interference contributions

$$\begin{aligned} \left[\left(\frac{C}{N}\right)_T\right]^{-1} &= \left[\left(\frac{C}{N}\right)_{UT}\right]^{-1} + \left[\left(\frac{C}{N}\right)_{DT}\right]^{-1} \\ &= \left[\left(\frac{C}{N}\right)_U\right]^{-1} + \left[\left(\frac{C}{I}\right)_U\right]^{-1} + \left[\left(\frac{C}{N}\right)_D\right]^{-1} + \left[\left(\frac{C}{I}\right)_D\right]^{-1}. \end{aligned} \quad (14.18)$$

Based on Eq. 14.18, an optimization of the major uplink and downlink parameters can be performed for the maximum C/N of the overall link. A general illustration of link optimization for multi-carrier operation through a common high power amplifier is shown in Fig. 14.15. As can be seen, an increase in the HPA back-off will improve the inter-modulation interference. At the same time, it will reduce the usable power to the downlink carrier, and therefore reduce the downlink C/N . As a result, the overall link C/N reaches its maximum around 2.5 to 3 dB output back-off for the amplifier.

In addition to the impacts from the thermal noise and other sources of interference, the imperfect satellite transponder frequency responses, group delay variations, and AM-AM and AM-PM nonlinearity of an HPA will distort the transmit signals and introduce additional errors to the communications link. The major components that contribute to the impairment of a typical communication satellite are the channelization filters in the input multiplexers, the filters in the output multiplexers, and the traveling wave tube amplifiers (TWTAs), as illustrated in Fig. 14.16.

To minimize the impact of those components on system performance, particularly in a single carrier per transponder operation condition as commonly used in DTH TV/data services, the impacts of the input multiplexer (IMUX) filter may be improved by applying proper pre-compensation techniques on the modulator [19]. As an example, Table 14.5 shows the transponder impairments and possible improvement with pre-compensation techniques for a DVB-S2 system at $BER = 10^{-5}$. As can be seen, the quasi-constant envelope modulations, such as QPSK and 8PSK, can operate on a transponder near saturation without significant system performance degradation, while 16APSK and 32APSK,

Table 14.5 Total C/N loss from saturated un-modulated carrier with and without pre-compensation

Modulation/ coding	Without pre- compensation	With pre-compensation
QPSK, $r = 1/2$	0.6 dB (OBO 0.3 dB)	0.5 dB (IBO 0 dB, OBO 0.4 dB)
8PSK, $r = 2/3$	1.0 dB (OBO 0.4 dB)	0.6 dB (IBO 0 dB, OBO 0.4 dB)
16APSK, $r = 3/4$	3.2 dB (OBO 1.7 dB)	1.5 dB (IBO 1 dB, OBO 1.1 dB)
32APSK, $r = 4/5$	6.2 dB (OBO 3.7 dB)	2.8 dB (IBO 3.6 dB, OBO 2 dB)

OBO output back-off, IBO input back-off

which are inherently more sensitive to nonlinear distortions and would require operation in quasilinear-transponders, derive the greatest benefit from the pre-compensation technique.

In FDM configurations, where multiple narrowband carriers occupy the same transponder, the transponder input and output multiplexer filters have only small phase variations (group delays) across each carrier spectrum that this impairment to system performance may be neglected. However, in this case the transponder must be kept in the quasilinear operating region (i.e. with large output back-off) in order to avoid excessive inter-modulation interference between signals. The impact of the inter-modulation can be considered as an AWGN interference and included directly in link budget computations.

14.4 Communications System Architectures

14.4.1 System Architectures for Communications Satellites

As shown in Fig. 14.2, a bent-pipe communications payload consists of an input (or receive) section and an output (or transmit) section. Communications systems for bent-pipe satellites are designed to provide their required functionality with sufficient performance and reliability over the mission life of the satellite. Typically, the overall satellite system reliability is specified to be greater than 0.8 after 15 years in orbit. This reliability is apportioned between the bus subsystems and the payload. A payload reliability greater than 90 % over a 15-year lifetime is usually specified. To achieve the payload reliability requirements, all failure-prone active components of the payload are provided with redundant units that can be switched in place of failed units. The functions of each of the payload elements and their impact on the overall performance of the system are discussed in this section.

14.4.1.1 The Input or Receive Section

The input or receive section consists of the receive antenna, the input filter, the receivers and the input multiplexer. The functions performed by the input section elements are:

1. Receive the uplink signals from the ground stations located in the receive coverage area. This function is performed by the receive antenna. Depending on the coverage area and polarization, the receive antenna could be a global horn, a shaped reflector antenna (including a dual gridded reflector antenna), or a multiple-feed spot beam reflector antenna. The receive antenna is followed by the input band pass filter that passes the desired receive band, while rejecting unwanted potentially interfering uplink signals as well as signals in the transmit band of the payload. This is essential to protect the receivers from the power that can leak from the transmit section of the payload. Typically the input filter provides about 80 dB of rejection to signals in the transmit band of the payload, while its insertion loss in the receive band is on the order of 0.2 dB, thus minimizing the impact on the receive gain to noise temperature ratio (G/T).
2. Low-noise-amplify and frequency translate the received signals to the downlink frequency bands. These functions are performed by the receivers. The receiver consists of a low noise amplifier followed by a down converter. For each coverage area and polarization there is a corresponding receiver. The receivers are configured to have redundancy, to achieve the desired reliability over the design life of the satellite, and to protect against possible on-orbit failures (Fig. 14.17).
3. Channelize the translated wideband signals into narrowband channels. This channelization is used to facilitate signal routing to different downlink beams and to allow separate transmit amplifiers to provide the downlink RF power for maximizing the DC-to-RF power conversion efficiency. The channelization is performed by the IMUX. The most common IMUX configuration is the channel-dropping configuration shown in Fig. 14.18. In this configuration, the wideband signal is split by a 3 dB hybrid. One output of the hybrid is connected to a group of filters corresponding to the even numbered channels, while the other output is connected to a group of filters corresponding to the odd numbered channels. The filters are connected through circulators, which provide the necessary directional isolation. Separating the channels by odd and even numbers in a channel-dropping configuration improves the in-band channel performances as compared to the case where the channels are contiguous.
4. Provide interconnectivity (switching) among the channels. This switching is required to direct channels to the appropriate intended coverage on the Earth.

Fig. 14.17 Receivers redundancy configurations

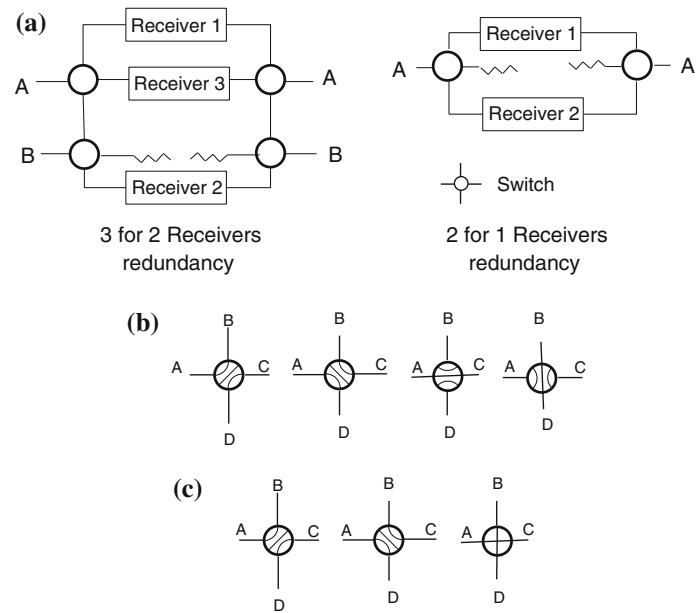
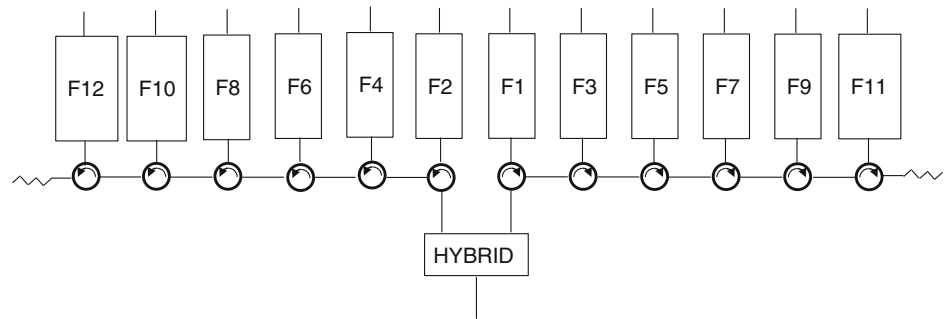


Fig. 14.18 An example of 12-channel input multiplexer



14.4.1.2 The Output or Transmit Section

The output or transmit section consists of transmit high power amplifiers including their input and output redundancy switching networks, the output multiplexer (OMUX), and the transmit antenna. Functions performed by the output section elements are

1. Amplify the signals in each channel (transponder) to provide sufficient transmit power. The power amplification is performed by high-power amplifiers, which are either TWTAs or solid-state power amplifiers (SSPA). The HPAs are configured to have redundancy to provide the desired reliability over the design lifetime, and protect against possible failures in orbit. The most common configuration is double-ring redundancy [20]. In this configuration there are a total of M power amplifiers, of which $N < M$ are active, and $(M - N)$ are in standby. The N inputs are connected by a group of switches and the M outputs of the amplifiers are selected by a group of interconnected switches. An example that illustrates the concept of ring redundancy is shown in Fig. 14.19, in which there are eight active amplifiers ($N = 8$) and four
2. Combine the amplified signals into one or more common output ports to be fed to the transmit antenna. This function is performed by the output multiplexer (OMUX). The OMUX for an N -channel payload has N low loss waveguide filters, connected to a waveguide manifold, with one end shorted and the other end constituting the common port to be connected to the transmit antenna. Waveguide filters and manifolds are used in the OMUX to provide the lowest possible loss, and to maximize the power handling capability. Figure 14.20 shows an example of an 8-channel OMUX. Even so, the output loss of the multiplexer is not insignificant and depends on channel bandwidth, frequency band, and the number of poles (resonators) in each channel filter. Typical values of OMUX loss using 4-pole filters and a

standby amplifiers ($M = 12$). The interconnections are achieved by R-switches and/or T-switches. The switches are connected to form rings on the input, and mirror images of these rings on the output. This topology allows operation of all eight channels with failures of any number from one to four out of the eight active amplifiers.

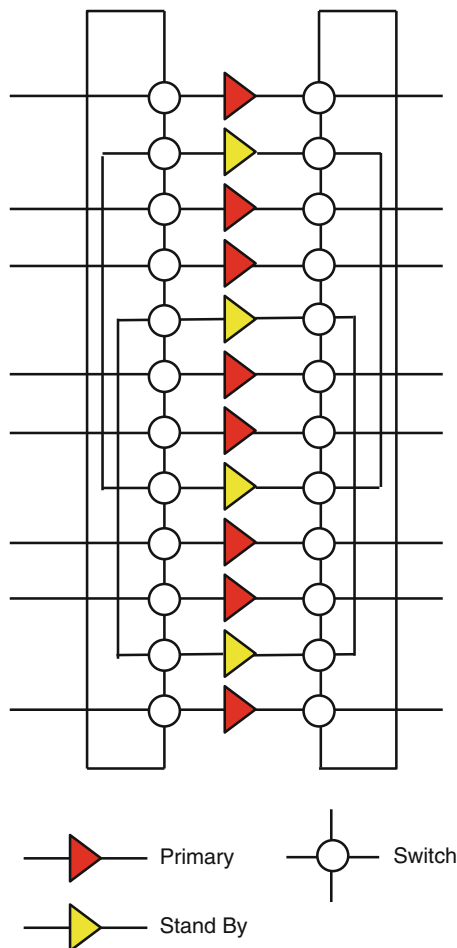


Fig. 14.19 Example of ring redundancy of traveling wave tube amplifiers (TWTAs)

36 MHz channel bandwidth are 0.25 dB in the C-band and 0.45 dB in the Ku-band.

3. Transmit the signals down to the required coverage areas on Earth. This function is performed by the transmit antenna.

Major factors to consider in the design of the communications payload are the antenna configurations, the HPA efficiency, reliability over the life of the satellite, and the total mass of the payload. The available payload power from the satellite platform and the payload mass that can be supported by the platform and launch vehicle directly define the maximum achievable communications capacity, i.e. the number of active operational transponders with the required EIRP. The typical communications satellite operational life is 15 years. Usually this life time is limited by the on-board fuel available to perform station-keeping maneuvers, by the progressive degradation of the solar panels, and to a lesser extent by the life expectancy of the payload active components. In the bent-pipe architecture, the active components are primarily the receivers and the HPAs. To achieve the required minimum lifetime the reliability of the payload

redundant (spare) units for all active components are employed. Properly configured switching networks that can select standby units in response to the failure of operating units provide this redundancy.

14.4.2 System Architectures for Telemetry, Tracking, and Command

Each spacecraft has to have a communications subsystem that provides the interface between the spacecraft and the ground control centers (or other relay satellites) for proper operation of the spacecraft. This subsystem is usually referred as the telemetry, tracking, and command (TT&C) subsystem. For most applications, this subsystem also provides the ranging capability and therefore is also referred as the telemetry, commands, and ranging (TC&R) subsystem. Through this subsystem, the mission control center can operate the spacecraft and its payloads based on the mission needs via ground commands, and receive the spacecraft housekeeping/health data as well as the mission data. Due to the criticality of this subsystem for a mission, redundancy/cross strapping of active flight units and near 4π (spherical) antenna coverage are essential to ensure a high reliability and a near-continuous communications link between the spacecraft and the ground control center for all mission phases/modes. Figure 14.21 shows a typical block diagram of a TT&C subsystem.

The command signal is typically a narrowband of the order of 1 MHz. The signal is first FSK or PSK modulated and then FM or PM modulated to a subcarrier. This modulation scheme is tolerant of the multipath effects introduced by the multiple command antennas and the spacecraft scattering due to the wide beamwidths of the command antennas. The redundant command receivers receive the command signal, low-noise amplify and down-convert it to an intermediate frequency (IF), demodulate the signal analogically and then digitally to form the command data stream. The demodulated commands are routed to the spacecraft control and data handling (C&DH) subsystem and then properly decrypted and executed via the C&DH subsystem and the flight computers. As indicated in Fig. 14.21, the redundant command receivers are cross strapped to the redundant processing units in the C&DH subsystem for high system reliability.

The spacecraft housekeeping/health data (as well as the mission data, for many science and technology and/or remote sensing spacecraft), are collected, formatted, encrypted, and stored (if necessary) by the C&DH subsystem and then distributed to the redundant telemetry transmitters for downlink to the control centers. The telemetry transmitter accepts the data from the C&DH subsystem, digitally modulates the data, and then analog modulates the signal to a subcarrier together with the possible ranging signals.

Fig. 14.20 An 8-channel output multiplexer using 4-pole dual-mode filters

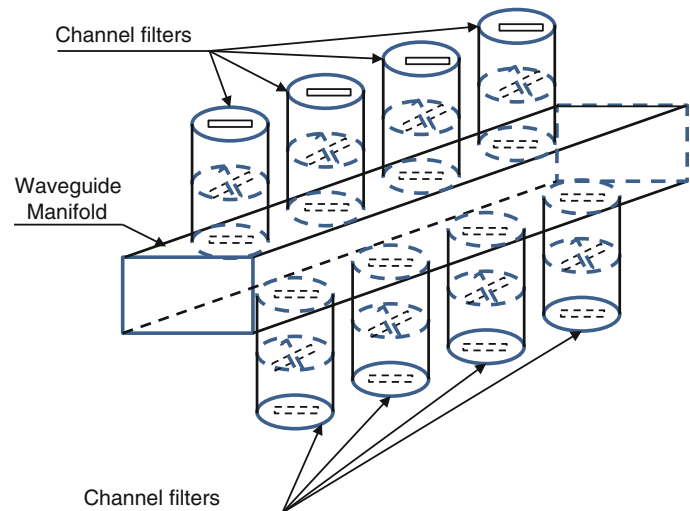
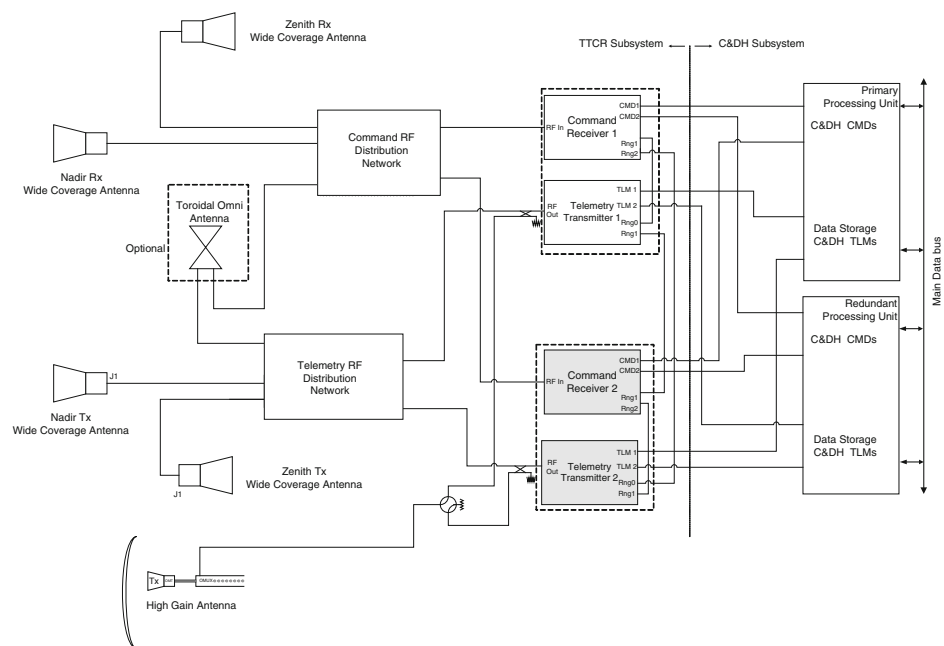


Fig. 14.21 A typical block diagram of a telemetry, tracking, and command (TT&C) subsystem



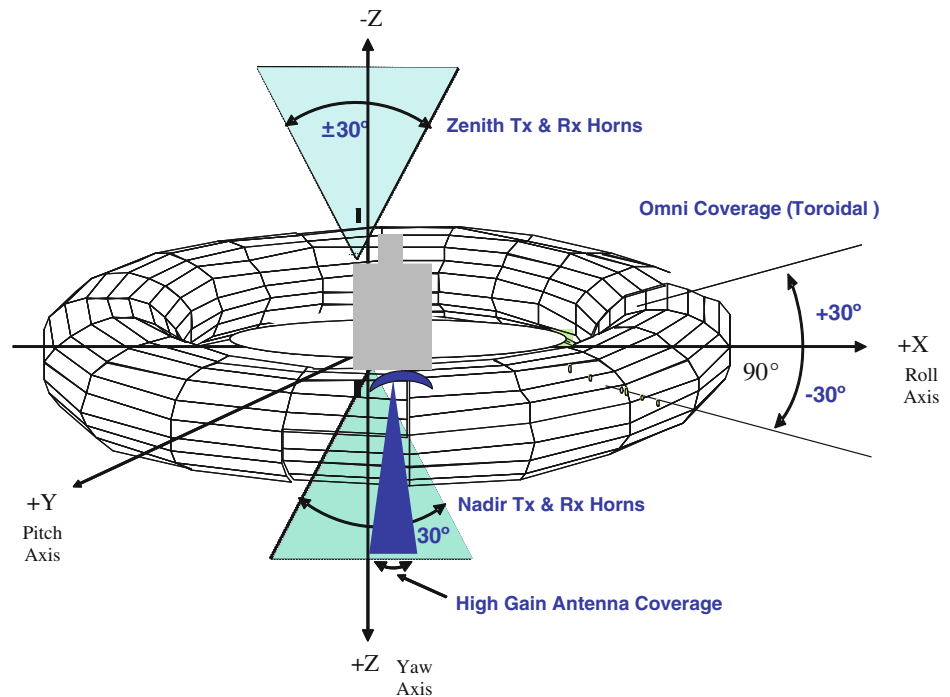
Multiple antennas are employed in the TT&C subsystem architecture, as shown in Fig. 14.22, to provide near-spherical composite coverage for receiving the command signals and for transmitting the telemetry signals, respectively. These antennas include the wide coverage antennas in the nadir direction and zenith direction as well as the omni antenna with a toroidal shaped radiation pattern. The high gain antenna is usually used for on-station nominal operation for a geosynchronous communications satellite, and for transmitting high-speed telemetry/mission data for a science and technology satellite or a remote sensing satellite.

For some types of spacecraft, the omni antenna with a toroidal shaped radiation pattern may not be required if the nadir and zenith wide coverage antennas are designed to provide near-hemispherical coverage.

14.4.3 System Architectures for Remote Sensing and Sciences/Technology Satellites

Unlike communications satellites, whose main function is to provide communications links to convey information between two points, or among multiple points, a remote sensing or science/technology satellite's main function is to collect information and/or images of objects of interest and transmit the collected information/images to ground collection centers for further study, processing, and analysis. When the data volume is limited, the mission data can be transmitted through the TT&C subsystem together with the spacecraft housekeeping/health data. However, when a large amount of mission data/images must be continuously

Fig. 14.22 Typical composite antenna coverage requirements for a TT&C subsystem



downlinked in real-time, a dedicated wideband downlink system will be required, as shown in Fig. 14.23.

14.4.4 Key Communications System Performance Analysis and Budget

Important parameters that define the communications performance include

1. Channel bandwidth and number of operating channels. Typical channel bandwidths in C-band and Ku-band payloads are 36, 54, or 72 MHz. The total available spectrum and the selected channel bandwidth together determine the total number of operating channels. For example, the available 500 MHz spectrum for each polarization allows for twelve channels (per polarization) of 36 MHz each, with nominal guard bands between adjacent channels of 4 MHz. For 54 or 72 MHz channels, the required guard bands are nominally 6 and 8 MHz respectively. For Ka-band spot beam systems, channel bandwidths of 125 or 250 MHz are common place. The channel bandwidth selection is a balance between spectrum efficiency usage and the available RF power in each channel. An example of a typical frequency channelization plan of a Ku-band payload is shown in Fig. 14.24. The payload has twelve 36 MHz bandwidth vertically polarized channels and eight horizontally polarized channels (four of 72 MHz bandwidth and four of 40 MHz bandwidth) on the uplink. The downlink has the same arrangement with orthogonal (horizontal) polarization.
2. Receive and transmit coverage areas. Satellite coverage areas can be global, shaped areas, or spot beams. Global beams cover the entire Earth visible from the satellite orbital location. Global coverage allows complete connectivity among locations on the visible Earth disk from the satellite position. However, a global beam allows only frequency reuse by polarization isolation. In addition, the maximum gain achievable by global beam are limited to approximately 18 dB. Shaped beams coverage is used to maximize the gain over the desired regions, while minimizing it over adjacent regions, thereby enabling frequency reuse by spatial isolation. Similarly, spot beams are utilized to maximize the gain over very small areas, for example large cities, while minimizing interference with other spot beams and allowing many frequency reuses. Spot beams are more often employed in mobile satellite service systems and wideband Ka-band systems.
3. Receive gain to noise temperature ratio (G/T). The satellite G/T ratio is an important measure of the signal degradation due to system noise. As illustrated in Sect. 14.3.1, G/T is primarily determined by the receive antenna gain, the noise figure or noise temperature of the low noise receiver, and the Earth-temperature background noise captured by the receive satellite antenna aperture.
4. Transmit equivalent isotropic radiated power (EIRP). This is an important measure of the downlink performance, and is expressed in dB Watt (dBW). The EIRP is the product of the transmitted power through the antenna times the transmit antenna gain as described in Sect. 14.3.1. EIRP is the major factor that determines the user

Fig. 14.23 A typical block diagram of a data downlink system

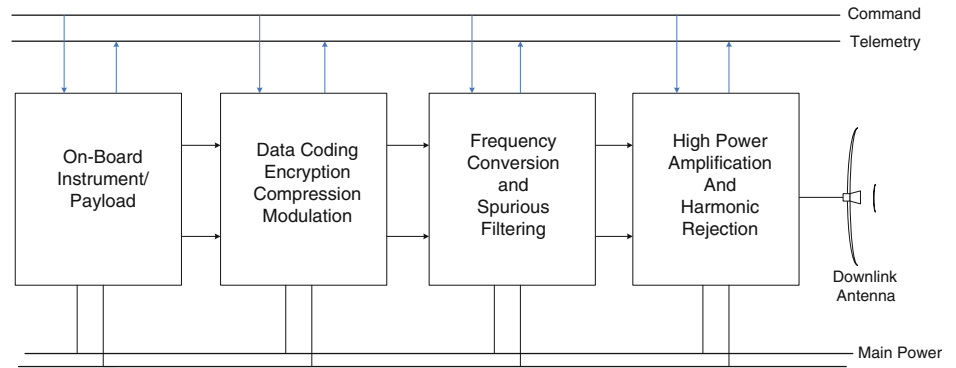
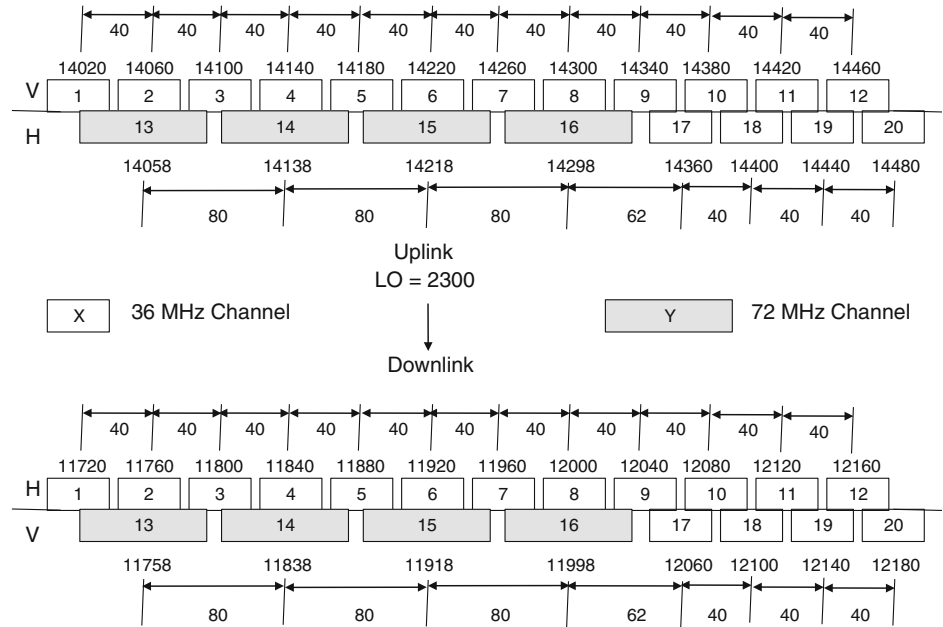


Fig. 14.24 Typical frequency plan of a Ku-band communications payload



terminal size on the ground for the reception of communications signals, such as DTH television broadcasting services.

5. Saturated power flux density (SFD). This is a measure of the input power density at the satellite’s receive antenna aperture that, given the transfer function of the satellite receive subsystem and channelization gains and losses, drives the high power downlink amplifier(s) to saturation. The saturation flux density is expressed in dB W/m².
6. Overall payload gain. This is defined by the EIRP and the SFD requirements.

Key payload system performance is determined by the performance of the individual elements that constitute the payload. System performance analysis is carried out by tracing the signal levels at the interfaces of the various payload components, and adding their gains or subtracting their losses in dB. To illustrate this process, consider a typical example of a C-band transponder with an antenna beam shaped to cover the continental U.S. (CONUS). The receive antenna gain at the edge of the coverage area is

assumed to be 27 dBi, and that of the transmit antenna gain is 28 dBi. Figure 14.25a and b show the gain/loss and signal levels for the input section and output section, respectively; the diagram shows the minimum and maximum levels for both the primary and redundant paths.

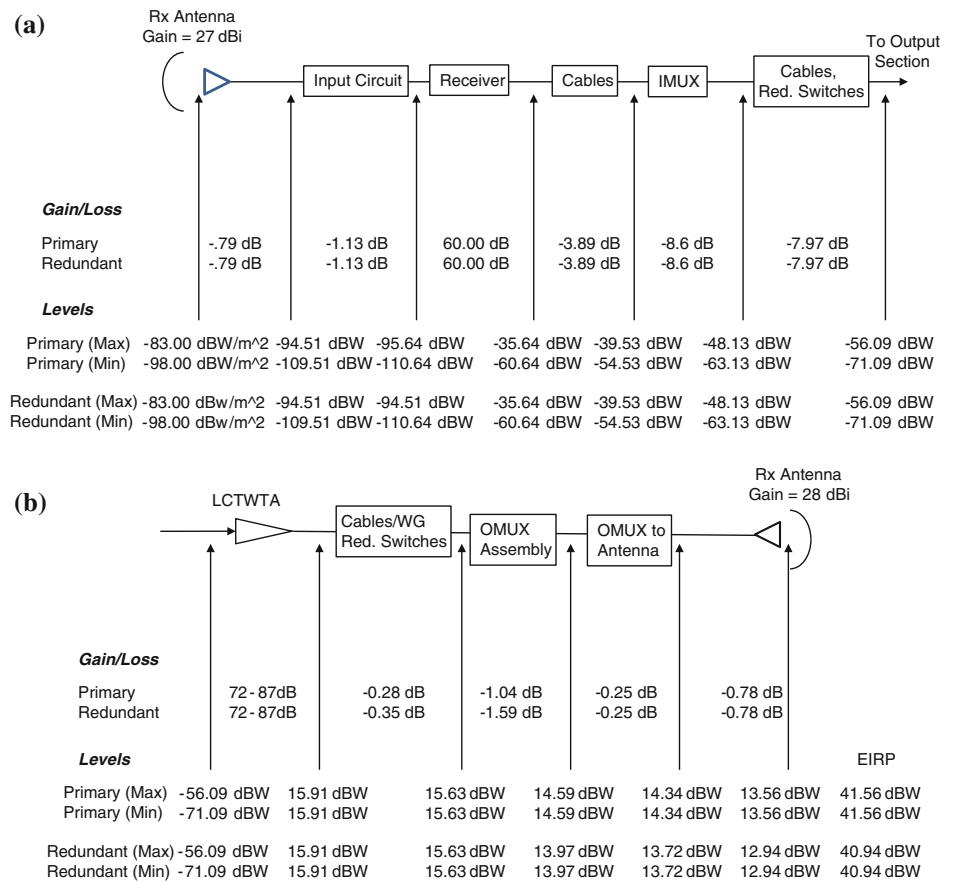
14.5 Communications Subsystem Characteristics and Performances

14.5.1 Antenna Subsystem

Antennas are used to radiate or receive the RF signal to or from space, respectively. They obey the reciprocity theorem and therefore all the properties of an antenna are the same in the transmit and receive applications. There are different types of antennas on spacecraft

- The omni-direction antenna and/or the wide beamwidth antennas used to receive and transmit telecommand and telemetry signals.

Fig. 14.25 Typical C-band payload gain/loss and signal Levels for **a** input and **b** output sections



- The high gain steerable spot beam antennas used on remote sensing and science/technology satellites for transmitting high data rate information back to Earth stations.
- The shaped beam coverage antennas and the multi-spot beam antennas for communications satellites.

The antenna subsystem is the transducer between free space (satellite) or air (Earth terminal) and the communications electronics. Its performance is tied by fundamental physics to its size, and unlike many circuit components it cannot be miniaturized. It is one of the most critical subsystems and often drives the spacecraft design and configuration. This is particularly true for a communications satellite where the antenna subsystem design must maximize the G/T and minimize the power required to meet the satellite EIRP. The antenna must also meet the cross-polarization isolation and the sidelobe isolation requirements. The isolation requirements enable frequency reuse to maximize the throughput of satellite communications with the limited spectrum available/assigned for satellite applications.

14.5.1.1 Antenna Fundamentals

An antenna is a component for conversion of an electrical signal into an electromagnetic wave or vice versa. The antenna radiation performance is characterized by its radiation pattern as shown in Fig. 14.26. Dependent on the

movement of the electric field vector of the radiated electromagnetic wave with time, an antenna can be linearly polarized if the direction of the vector of radiated electric field is constant with respect to time, or circularly polarized if the direction of the electric field vector rotates with respect to time and maintains a constant amplitude with that time rotation. The radiation pattern in its principal (desired) polarization is usually referred as the co-pol pattern and in the orthogonal polarization is referred as cross-pol (or x-pol) pattern. The ratio of the cross-pol pattern to the co-pol pattern is defined as cross-pol discrimination (XPD).

A fundamental parameter of an antenna performance is its directivity D , which measures the radiation intensity (i.e. radiated power/solid angle) in a specific direction and a particular polarization to the average radiation intensity over all directions (a sphere). The directivity depends only on the shape of the radiation pattern and can be given by

$$D(\theta, \varphi) = \frac{P(\theta, \varphi)}{P_{average}} = \frac{P(\theta, \varphi)}{\frac{P_r}{4\pi}} = \frac{P(\theta, \varphi)}{\frac{1}{4\pi} \int_0^{2\pi} \int_0^\pi P(\vartheta, \psi) \sin \vartheta d\vartheta d\psi} \tag{14.19}$$

where $P(\theta, \varphi)$ is the radiation intensity (power pattern function) of the antenna in a particular direction and polarization, $P_{average} = P_r/4\pi$ is the radiation intensity

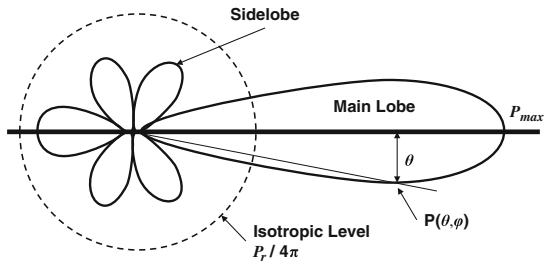


Fig. 14.26 Antenna radiation pattern

averaged over all space, and P_r is the total power radiated by the antenna.

A closely related parameter is the antenna gain G , which is the ratio of the radiation intensity in a specific direction and polarization to the radiation intensity that would be produced by an isotropic radiator accepting the same input power P_{in} , that is

$$G(\theta, \varphi) = \frac{P(\theta, \varphi)}{P_{in}/4\pi} = D(\theta, \varphi) \cdot \frac{P_r}{P_{in}}. \quad (14.20)$$

Thus, the antenna gain accounts for the ohmic losses in the antenna. While not included in the IEEE definition of antenna gain [21], the reflection loss due to mismatching impedances must be included in the link budgets. For a lossless antenna, the antenna directivity will be equal to the antenna gain.

The peak directivity D_p of an aperture antenna is the directivity in the maximum radiation direction and can be calculated as

$$D_p = \frac{4\pi A_e}{\lambda^2} = \frac{4\pi}{\Omega} \quad (14.21)$$

where λ is the wavelength of the radio wave, A_e is the effective aperture area; and Ω is the angular area (or solid angle) within which the antenna focuses the power. In general, the effective aperture area is less than the physical aperture area due to nonuniform field distribution over the antenna physical aperture. The usefulness of the effective aperture area is in its ‘collecting’ function: for a particular incident power flux density (PFD), the received power is $P_R(W) = PFD(W/m^2) \cdot A_e(m^2)$.

Satellite antennas generally do not try to maximize the peak directivity or gain. Rather, the design objective is to maximize the minimum antenna directivity or gain over a finite angular region (solid angle) of space, e.g. to cover a country or specific region on the Earth or oceans. So-called shaped beam antennas are fundamentally limited by their size in the degree to which they can arbitrarily increase the minimum gain or directivity over a finite given angular region. To concentrate all its energy into a given non-zero angular area of Ω steradians in order to achieve the

coverage directivity of $4\pi/\Omega$ as defined in Eq. 14.21, an infinitely large antenna would be needed. Therefore, $4\pi/\Omega$ represents the maximum achievable coverage directivity for a coverage with an angular area of Ω .

The coverage area of a shaped beam is usually defined by multiple coverage polygons, and each polygon usually requires a different coverage directivity. As a useful concept and tool for coverage performance estimates, the maximum achievable directivity can be expanded to each coverage polygon in a beam that has multiple coverage polygons, as shown in Fig. 14.27. Here, polygon 0 is the main polygon and the maximum achievable directivity is D_0 , polygon 1 to polygon n are the congruent polygons (i.e. polygons contained within main polygon p_0), polygon $(n + 1)$ to polygon $(n + m)$ are the non-congruent polygons, and Ω_i is the solid angle of each polygon. The maximum achievable directivity of each polygon is normalized to D_0 as $\alpha_i = D_i/D_0$.

Assuming that a hypothetical ideal antenna would produce a constant step directivity over each polygon, the maximum achievable directivity of the main polygon can be derived as

$$D_0 = \frac{4\pi}{[\Omega_0 + \sum_{i=1}^n (\alpha_i - 1)\Omega_i + \sum_{i=n+1}^{n+m} \alpha_i \Omega_i]}. \quad (14.22)$$

In practical applications, the realizable directivity over each polygon is typically around 35–70 % of the ideal directivity, depending on the coverage area and its shape, due to the finite antenna size, the spillover loss, and other factors.

14.5.1.2 Satellite Antenna Technologies and Implementations

Antennas used for space communications can generally be divided into low or medium gain antennas (gain from 0 to 20 dBi) mainly used for TT&C applications, Earth coverage applications, radiation elements for phased array antennas, and primary feeds for reflector antennas, and high gain antennas (gain >20 dBi) for high data rate communications to a regional coverage or a spot coverage.

The commonly used low or medium gain antennas include biconical antennas, helical antennas, dipole or patch excited cup antennas, and horn antennas. Figure 14.28 shows pictures of a horn antenna and a biconical antenna, as well as their radiation patterns. A summary of low and medium gain antennas for small satellite applications can be found in [22].

High gain antennas require large effective radiation apertures, which can be provided by reflectors, arrays of radiating elements with low/medium gains, and/or lenses.

The reflector antenna is the most commonly used antenna configuration in satellite communications because of its simplicity and light weight [23, 24]. It consists of reflector with a feed or an array of feeds located in or near the focal

Fig. 14.27 Example of antenna coverage definition with multiple polygons

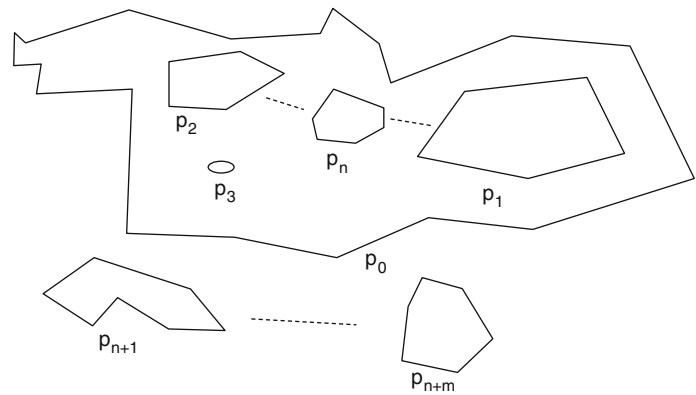
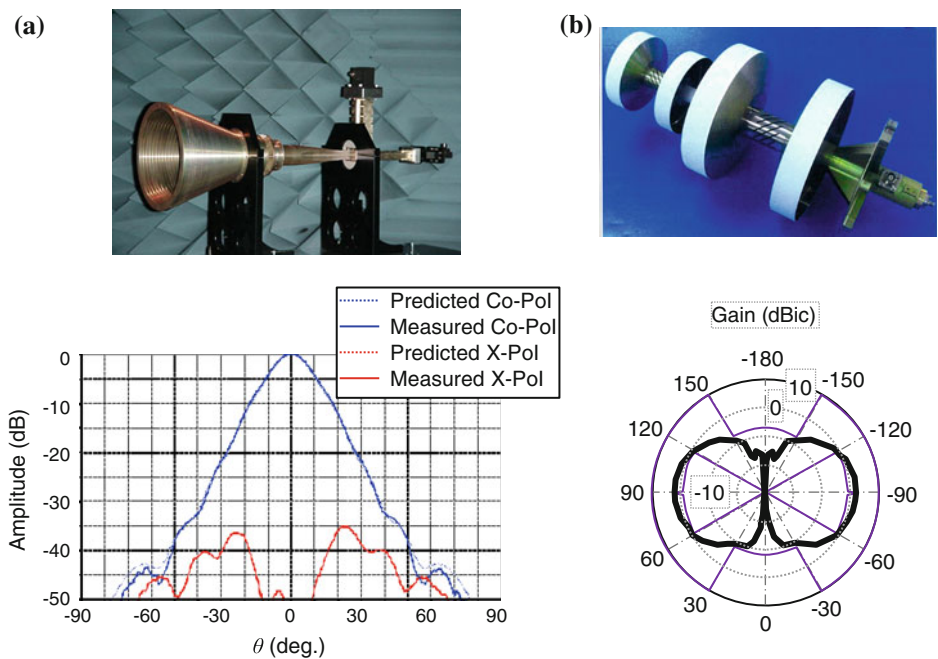


Fig. 14.28 **a** Horn antenna and typical radiation pattern; **b** biconical antenna and typical radiation pattern. *Image Orbital (left) and Rymsa (right)*



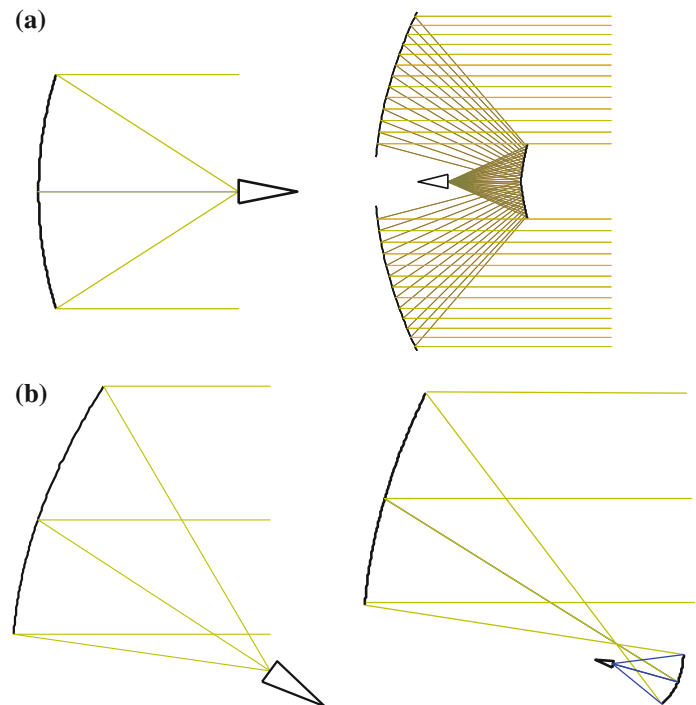
point. The reflector can be in an axially symmetric configuration that forms a centrally fed reflector antenna, as illustrated in Fig. 14.29a. In practice, the focal length of a reflector antenna usually is limited, resulting in a degradation of cross-pol performance due to finite curvature of the reflector and the pattern degradation for the feeds that are not located at the focal point. To minimize the possible performance degradations, a dual-reflector configuration can be used where the main reflector is illuminated by a combination of a primary feed(s) and a sub-reflector using portions of a hyperbolic surface (Cassegrain configuration) or ellipsoidal surface (Gregorian configuration). The radiation pattern of an axially symmetric configuration suffers from the blockage effect of the feed(s) and/or the sub-reflector, resulting in peak gain reduction and the sidelobe level increase. In addition, accommodation of a centrally fed reflector antenna with a large main reflector in a launch vehicle is usually found to be challenging due to finite fairing dimensions.

The aperture blockage can be avoided by using either a single offset reflector configuration or a dual-offset reflector configuration, as shown in Fig. 14.29b. Compared to the centrally fed reflector configurations, the offset reflector configurations can be more easily accommodated in the fairing of a launch vehicle since the main reflector can be readily made deployable from a side of the spacecraft.

To further improve antenna cross-polarization performance and to allow one reflector antenna to provide multiple coverage beams in different polarizations and/or different frequency bands to minimize the spacecraft mass and to ease the launch vehicle accommodation, polarization selective surface and/or frequency selective surface (dichroic surface) may be used for the main and/or the sub-reflectors.

Figure 14.30 shows a dual-gridded reflector (DGR) antenna configuration that has been widely used in communications satellites to improve the cross-pol performances of linearly polarized antennas. The antenna consists of two

Fig. 14.29 Commonly used reflector antennas for **a** axially symmetrical configuration and **b** offset configuration



reflectors with different focal lengths. The front reflector is a polarization selective surface. It consists of conducting grids that are parallel to the polarization direction of the signal to be reflected, and a supporting shell that is constructed with RF transparent materials. The front reflector will reflect the signal that has its vector of electric field parallel to the grid direction and will allow the signal in the orthogonal polarization to penetrate the shell with minimum loss. The rear shell can be a solid reflecting surface or a surface with grids in the orthogonal direction. The signal in the orthogonal polarization that penetrates the front shell will be reflected by the rear surface. The two surfaces can provide either congruent beams or divergent beams. Due to the filtering effect of the grids, the antenna cross-pol performance can be significantly improved (>10 dB improvement) compared to the conventional single-offset reflector antenna configuration.

Combining polarization selective surfaces and frequency selective surfaces, reflector antennas can be designed to operate in multiple frequency bands and polarizations. Figure 14.31 shows an example of a reflector antenna using a frequency selective surface that reflects Ka-band (30/20 GHz) signals but is transparent to Ku-band (14/12 GHz) signals, as the sub-reflector of a dual-reflector configuration for the Ka-band. In the Ku-band, the antenna is operated in a single-offset configuration and the polarization selective surface separates the orthogonally polarized signals.

Satellites used for mobile satellite service (MSS) operate in the L and S-bands. These require small coverage cells on Earth with high satellite antenna directivity in order to enable multi-fold frequency reuse and the use of hand-held

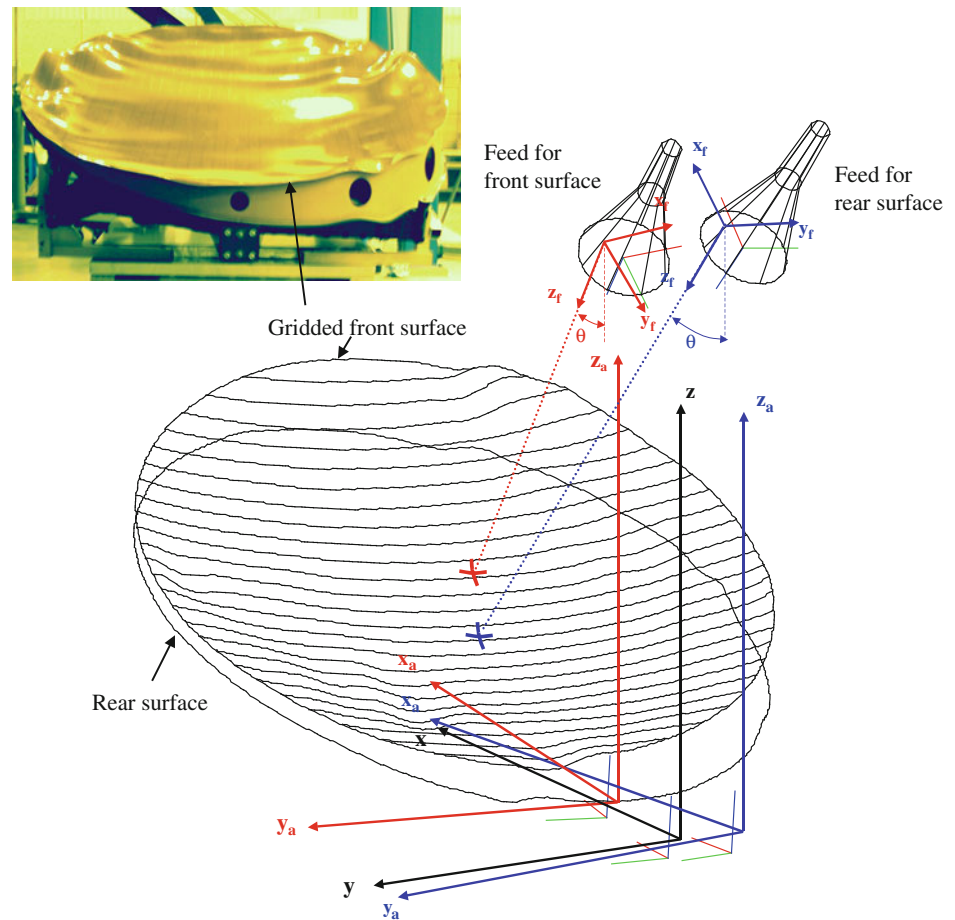
units for communications. The antenna reflectors required for this type of satellites are usually the unfurlable reflector antennas with the reflector size in 5 m to 30 m range, as shown in Fig. 14.32. Unfurlable reflectors exist with flight heritage in frequencies from UHF to the Ka-band.

In addition to reflector antennas, lens antennas have found applications in space communications. The principle of a lens antenna is illustrated in Fig. 14.33. Similar to a reflector antenna, which transforms a spherical wave from the primary source to a plane wave through reflection from a parabolic surface, a lens antenna achieves the same transformation via the refraction of the lens. The advantage of a lens antenna over a centrally fed reflector antenna is that a lens antenna completely avoids the blockage issue since the primary source (feed or feeds) is located behind the main radiation aperture. In comparison to offset reflector antennas, lens antennas are axially symmetrical and therefore provide better cross-polarization performance.

A lens antenna can be designed in many ways [24]. The lens formed by a homogeneous dielectric material provides wideband frequency performance, but is usually heavy. A waveguide lens with different zones is mass efficient but suffers from narrow frequency bandwidth. A comprise is a Bootlace (TEM) lens, which uses pick-up and radiating elements to receive and radiate the signals from the primary source and TEM transmission lines to provide the phase transformation from a spherical wave at the pick-up elements to a plane wave at the radiating elements.

Since the thickness of a lens is dependent on the wavelength of the operating signal, lens antennas are usually

Fig. 14.30 A dual-gridded reflector antenna. *Image* Orbital sciences corporation



bulky and heavy at low frequencies and are suitable only for some applications above 10–15 GHz.

A promising antenna technology for space communications is the phased array antenna technology. In a phased array antenna, multiple beams can be formed and electrically/electronically steered by feeding each radiating element of the array with a signal having certain phase and amplitude relationships with the others. The correct phase and amplitude for each element are generated by a beam forming network (BFN) that can be implemented using either analog or digital techniques. When the phase/amplitude relationships of the radiating elements are controlled by on-board and/or ground commands, a phase array antenna can provide the flexibility of dynamic beam reforming/reshaping and/or repointing/switching to meet the changes of communications traffic. It can also generate nulls at given directions to avoid harmful interference/jamming signals.

A phased array antenna may be formed by direct radiation elements or by a reflector/lens with an array of feeds. The former has been widely used in MSS low and medium Earth orbits, such as the Iridium and GlobalStar satellites, due to the relatively wide beamwidths required for those orbits. Geostationary MSS satellites such as NStar C,

Inmarsat, and ICO G1, and wideband satellites such as the Spaceway satellites, have applied the latter, for which a large deployable mesh reflector (5–30 m) antenna with a feed array is the preferred configuration in order to form small beams for high EIRP and G/T as well as to achieve a high degree of frequency reuse for high system throughput.

14.5.1.3 Shaped Beam Antennas and Spot Beam Antennas

For a communications satellite, the coverage areas (footprints), which are usually defined by the coverage polygons, determine the addressable market and the flexibility of extending services. To use the satellite power in the most efficient way, and to obey the coordination agreements among satellite service providers, it is necessary to design the satellite antennas such that their radiation patterns follow the coverage footprints. The antenna which provides the beam shape conforming to the coverage polygon is referred to as a shaped beam antenna. Figure 14.34 illustrates a typical coverage polygon definition for CONUS and the corresponding directivity contours of a shaped antenna. The most successful approaches for shaped beam antennas to date are (a) a standard (parabolic) reflector with a multiple feed array located in the vicinity of the focal point in

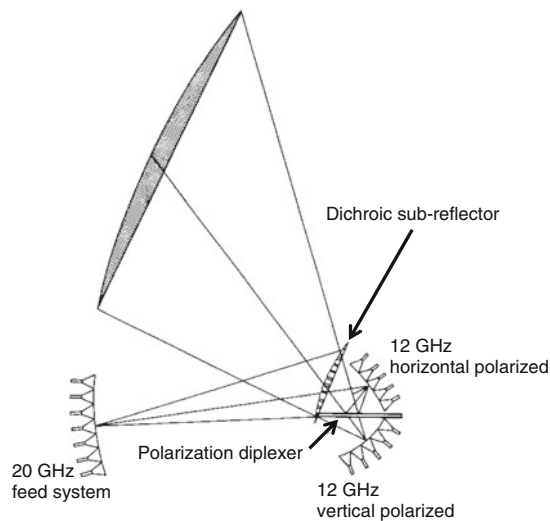


Fig. 14.31 Dual-frequency and polarization multi beam reflector antenna

conjunction with an associated beam forming network, and (b) a shaped surface reflector with a simple feed assembly. As illustrated in Fig. 14.35, the former realizes the shaped beam by combining the secondary pattern of each feed through the beam forming network, providing the required phase and amplitude excitations to each feed, while the latter realizes the beam shaping by deforming the reflector surface and thereby effectively changing the aperture field distribution.

For a given reflector size, both approaches will provide similar antenna directivity performance. However, since using a feed array involves a beam-forming network and this can become very complicated when a large number of the feed horns are used, the actual realized gain performance for the feed array approach will be less than the shaped reflector approach, due to increased ohmic losses. On the other hand, when an antenna is required to provide multiple shaped beams, such as the multiple cell coverage used by S-band MSS satellites like Thuraya and Inmarsat-4 and the multi-coverage C-band and Ku-band satellites like Intelsat V [15], or to allow on-orbit reconfiguration of the beam shape, the approach using a feed array can be more advantageous. In addition, the shaped beam using a feed array approach tends to provide better sidelobe performance and faster roll off due to the use of ‘whole reflector aperture’ for shaping [25].

The shaped beam over the entire coverage area allows the same signal to be delivered anywhere inside the coverage beam as shown in Fig. 14.34. However, this approach limits the overall satellite throughput mainly due to the limited use of available bandwidth as well as the effective spacecraft antenna gain (and hence EIRP and G/T) at the boundary. Two major approaches utilized to improve the bandwidth usage are to go to spot beam satellites and to use

adaptive coding and modulation (ACM) techniques as discussed in Sect. 14.3.2.

Spot beam satellites allow frequency reuse for spatially isolated beams so that a single satellite in orbit can have a large communications throughput, which lowers the bandwidth costs. There can be many frequency reuse schemes for contiguous spot beam configurations depending on the co-channel beam (i.e. beams using the same frequency) interference requirement. The commonly utilized frequency reuse schemes are shown in Fig. 14.36 [26, 27]. The set of contiguous beams that share the total available bandwidth is known as a cluster. The clusters are then repeated in the coverage area by relying on the fact that the beams operating at the same bandwidth will be separated from each other sufficiently to minimize their mutual interference. For a satellite with N spot beams and using an M color frequency reuse scheme, the satellite realizes (N/M) -fold frequency reuse. Reflector antennas with multiple feeds providing ‘single feed per beam’ configuration, phased array antennas, or multiple beam lens antennas can all provide spot beams. The multi-spot beam reflector antenna is commonly used in broadband satellite systems and the phased array antenna is more frequently adopted by MSS satellites.

Due to frequency reuse, co-channel beam interference generated by the surrounding beams using the same frequency channel becomes critical. Figure 14.37 illustrates the construction of interference within a contiguous spot beam configuration with three color frequency reuse. The overall co-channel beam interference is dependent on the spatial separation and the total number of the beams sharing the same frequency, but is dominated by the sidelobe response of the closest beams that share the same channel. The beam isolation requirement depends on the susceptibility of the modulation to co-channel interference and the dynamic range of the users. In general, the higher the number of colors in reuse, the lower is the co-channel beam interference.

A design conflict exists between the desire for beams with less spillover loss and low sidelobes to provide required beam isolation, and the desire for the beams to crossover at a high pattern level to maximize the minimum gain within the coverage area. Reduced spillover loss and sidelobe levels require an aperture distribution with low amplitude tapering that, in turn, requires directive antenna feeds. However, the antenna feed size is limited by the finite spacing between the beams using the same channel and the inability to physically overlap the antenna feeds. One approach commonly used in multi-spot beam broadband satellites to overcome the conflict is to utilize a minimum of three or four reflector antennas with each providing a set of multiple interleaved spot beams among the required contiguous beams as shown in Fig. 14.38 [27]. The scan distortion can be further reduced by increasing the focal length (or reducing the offset) or by using the dual-reflector

Fig. 14.32 TerreStar Networks, Inc.'s geostationary satellite, TerreStar-1 launched on July 1, 2009, aboard an Ariane 5 heavy-lift launch vehicle; shown at the launch facility (*left*). At the time it was the largest commercial satellite ever launched, carrying a state-of-the-art mobile satellite service (MSS) payload featuring a large 18 m unfurlable reflector built by the Harris Corporation; shown on the ground (*top, right*) and in orbit (*bottom, right*). *Image* Harris Corporation (*top, right*) and Space Systems/Loral (*left and bottom, right*)

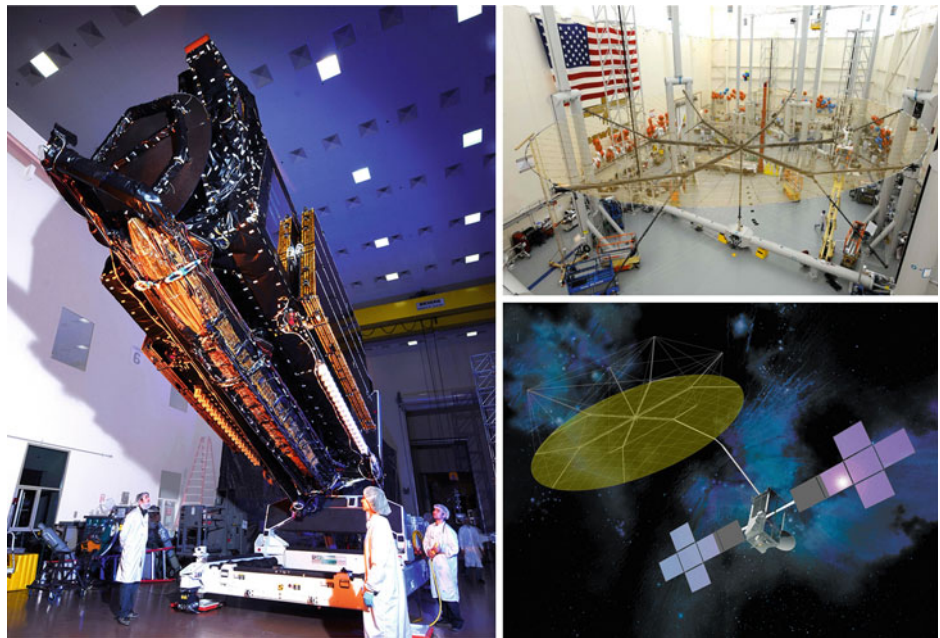
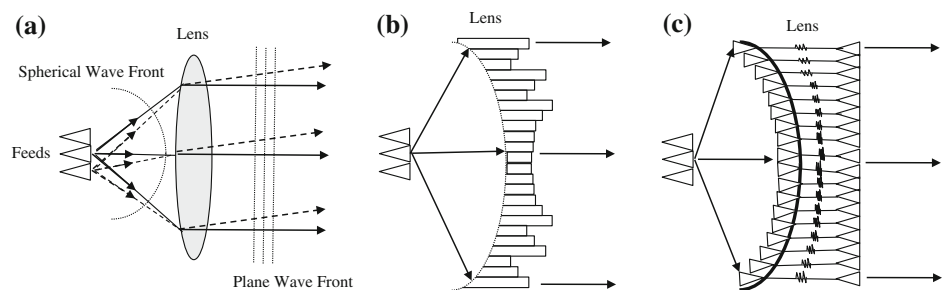


Fig. 14.33 Lens antennas for **a** dielectric lens; **b** waveguide lens; and **c** bootlace (TEM) lens



configuration (such as the Gregorian antenna configuration or the Cassegrain antenna configuration). Figure 14.39 depicts the typical edge of coverage (EOC) and sidelobe degradations with increasing scan angles of the defocused beams for different type of multi-spot beam reflector antennas [28].

14.5.1.4 Practical Considerations and the Impacts on Performance

In practice, the antenna design on a spacecraft will be constrained and impacted by multiple factors as summarized in Table 14.6.

The impacts of the factors in Table 14.6 on the antenna main-lobe performance, i.e. the edge of coverage (EOC) gain, are usually small for a properly selected antenna configuration and a well-designed antenna, and can be bounded by itemized antenna loss budgets as shown in Table 14.7. However, the impacts on low level co-pol and cross-pol patterns of an antenna can be significant, depending on the specific design and the required co-pol and cross-pol levels. Figure 14.40 shows a typical performance impact

(degradation) of the measured results with respect to the design on the low level co-pol and cross-pol for a Ku-band antenna in a flight configuration. A careful analysis and simulation of the antenna performance in flight configuration with accurate modeling techniques is necessary when low-level performance requirements become important.

14.5.2 Input Filter Assembly

The function of the input filter assembly is to reject the unwanted uplink signals, prevent the transmit signals from leaking to the receiver input in order to protect the receivers from the transmit signals, and to minimize the receiver intermodulation products among the uplink signals. The assembly consists of a cascade of a low pass filter and a band pass filter. Since this assembly precedes the receiver, it must have the lowest possible loss so that it does not degrade the input noise figure significantly, and hence it is implemented as a waveguide assembly. The band pass filter provides the necessary rejection of the unwanted signals

Fig. 14.34 A typical coverage directivity polygon over CONUS and the associated shaped beam contours

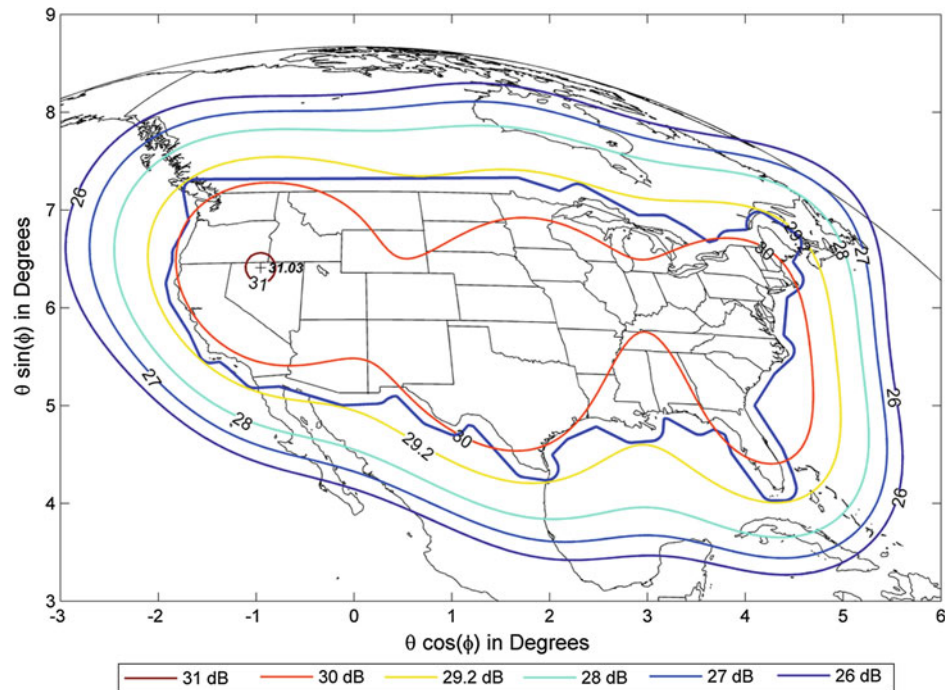
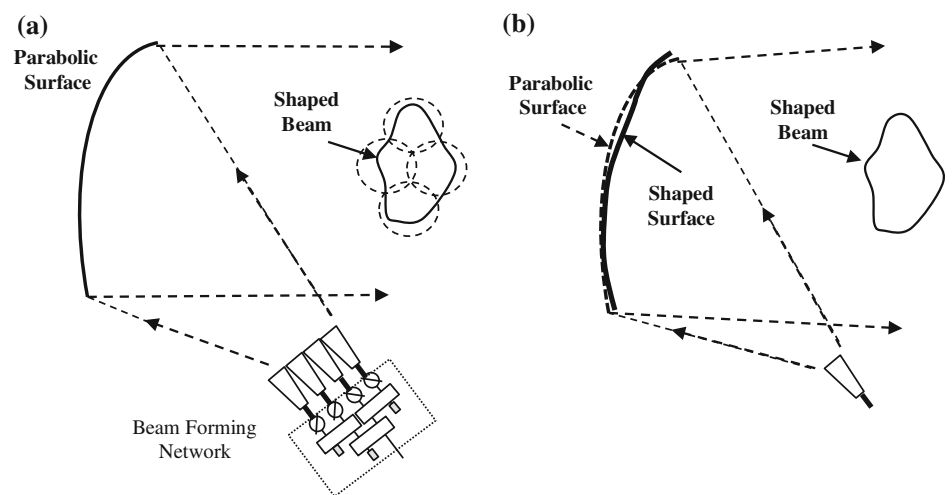


Fig. 14.35 Shaped beam antenna **a** implementation reflector antenna and a multiple feed array; and **b** implementation with a shaped reflector



close by, but outside the receive band, including the ring around transmit signal. As the waveguide band pass filter has spurious pass bands at the higher frequencies, the low pass filter is needed to provide the necessary rejections at the higher spurious bands, such as the harmonics of the downlink signals. Figure 14.41a shows a typical C-band input filter assembly, whose schematic diagram is shown in Fig. 14.41b. The input filter assembly must provide about 80 dB of rejection for the transmit signals, and for signals in the C-band as well as the Ku-band to protect the sensitive receivers, while the pass band loss should be less than 0.2 dB, in order to minimize the impact on the overall noise figure.

14.5.3 Low Noise Amplifiers and Receivers

The receiver assembly performs the functions of low noise amplification of the uplink signals, converting the signal frequency from the uplink band to the downlink band, and maintains linear operation across the input drive range. The low noise amplifiers are followed by a down converter. The down converter is composed of a mixer, a local oscillator, and filters to suppress the mixing products. The receiver has its own DC power section that provides the necessary power conditioning in order to minimize impacts of temperature variations on the stability of the overall receiver response. Figure 14.42 is a functional block diagram of the receiver,

Fig. 14.36 Frequency reuse schemes for **a** three colors (channels); **b** four colors (channels); **c** seven colors (channels)

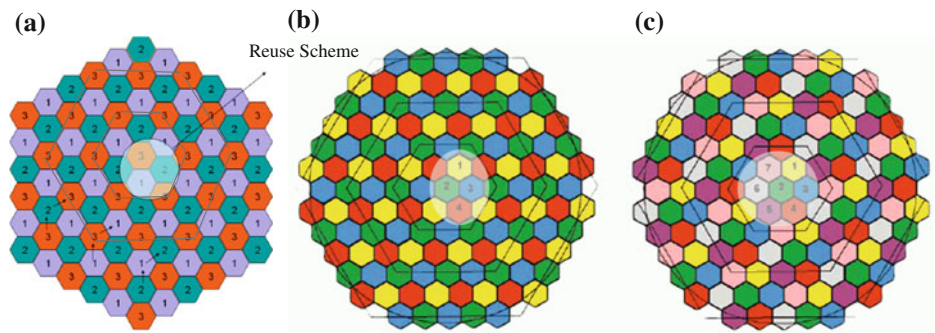
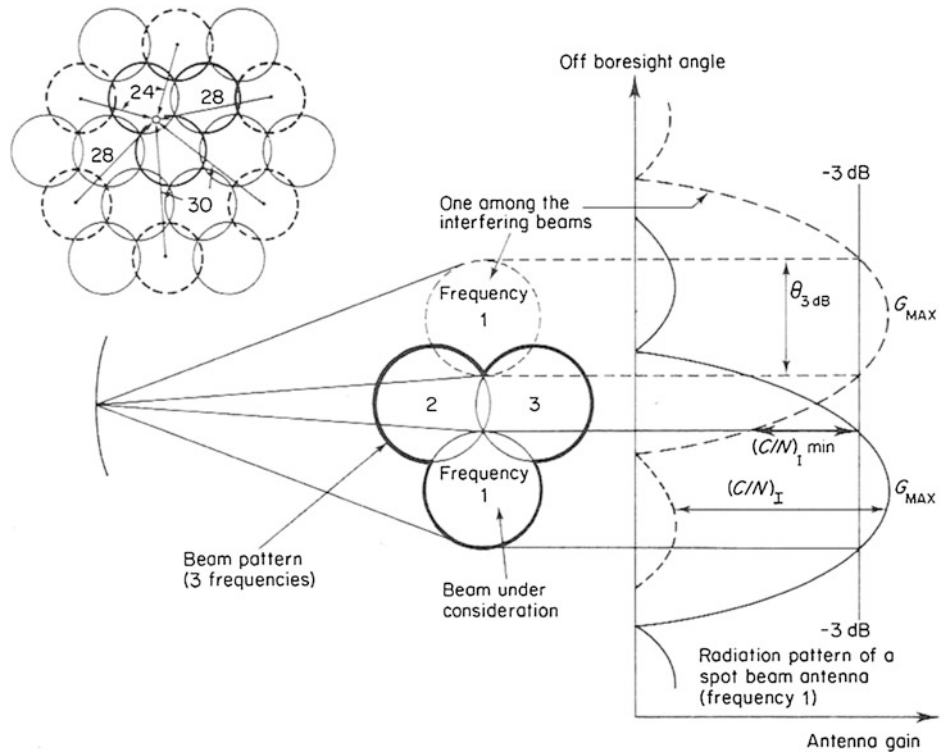


Fig. 14.37 Construction of co-channel beam interference within a beam lattice with three color frequency reuse scheme [15]



showing its three modules comprising an RF module containing the complete RF section and bias circuits, a local oscillator (LO) module, and a DC/DC converter module.

The important parameters for the receiver performance that affect the overall communications performance of the system and their nominal ranges of values are

1. The receiver noise figure. Typical ranges for the values of the noise figure are 1.4–1.6 dB for the C-band, and 1.8–2.0 dB for the Ku-band.
2. Gain and gain stability. Usually the receiver gain is in the range of 60 dB. Gain stability over temperature variations and over the lifetime is typically 1–2 dB. The receiver gain variation over the frequency band (gain flatness) is required to be within 2 dB.
3. Linearity. This is usually specified in terms of the carrier to third-order intermodulation C/I ratio when two low level equal carriers are present at the receiver input. A

typical C/I value greater than 42 dB is required when two equal carriers each at -47 dBm are present at the receiver input.

4. In-band spurious performance. This is the level of spurious intermodulation and harmonic signals from the LO that fall within the communications band. These levels are usually specified to be 60–70 dB below the input signal.
5. Out-of-band spurious performance. These are the levels of intermodulation products that fall outside the communications band, and are usually specified to be better than -60 to -65 dB below the input signal.
6. Phase noise spectral density. The LO phase noise spectral density is usually specified as a function of frequency offset from the LO frequency. Figure 14.43 is a typical specification of the LO phase noise variation versus frequency offset from the LO frequency.

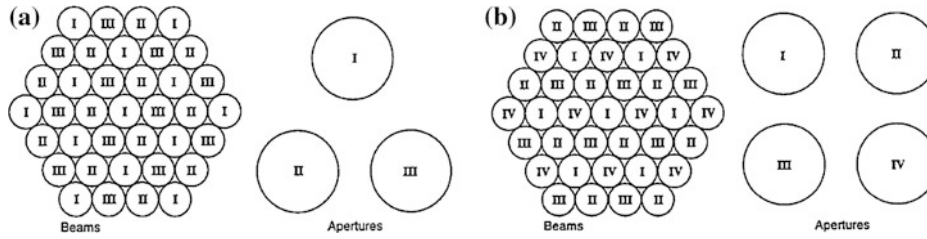


Fig. 14.38 Beam-aperture layout of multi-spot beam antennas using **a** three antenna apertures and **b** four antenna apertures

Fig. 14.39 Scanned beam performances of different types of reflector antennas [28]

Antenna Design	Antenna Geometry	Antenna Trade
Single-Offset Reflector		Simple to package but scan performance and sidelobe levels poor.
Offset Reflector		Folded optics provide longer equivalent focal lengths and improve scan loss.
Offset Gregorian		
Top-fed Cassegrain		Top-fed Cassegrain large F/D ratio and gives excellent scan performance
Side-fed Cassegrain		Side-fed Cassegrain also has excellent scan performance. Mechanical packaging is more compact for minimum stowage.

Gain reduction	Sidelobe level

Figure 14.44 shows a photograph of a flight C-band receiver.

14.5.4 Input Multiplexers

The functions of the input multiplexer (IMUX) are

1. To channelize the wideband signals into narrow channels, each of which will be amplified individually by a power amplifier (either a TWTA or a SSPA).
2. To reject spurious signals that are generated by the receiver.
3. To prevent the command signal from leaking through to the adjacent channels.

The important performance parameters of the IMUX are

1. Adjacent channel rejection. Each of the channel filters of the IMUX must have sufficient selectivity to prevent

interference from the adjacent channels. As was discussed in Sect. 14.4.4, the allocated frequency spectrum is divided into a number of channels with nominal bandwidths ranging from 27 to 125 MHz each. These channels are separated by guard bands, which are typically about 10 % of the channel bandwidths. The guard bands are required to allow practical filters to provide sufficient rejection of the adjacent channels. The IMUX channel filters are specified to reject an adjacent channel’s frequency with at least 15–20 dB at their band edges and 40–45 dB at the band centers. Figure 14.45a shows the frequency response including the near-band rejection of a typical IMUX filter of 36 MHz bandwidth. The required rejection specifications determine the number of filter sections (or poles) needed.

2. Insertion loss and group delay flatness. The pass band of each channel of the IMUX must introduce minimal

Table 14.6 Major design constraints and their impacts on spacecraft antennas

Item	Design drive	Impacts on design and performance
Launch vehicle	• Stowed configuration and size	• Antenna type and size – Realizable gain, sidelobe level and cross-pol level
	• Mechanical environments (vibration, shock, and acoustic)	• Light weight material (composite or mesh grids) – Reflector hydroscopic effect – On-orbit thermal distortion
	• Launch mass	– Depolarization – RF loss • Antenna mechanical structure – On-orbit thermal distortion – Scattering
Spacecraft accommodation constraints	• Overall layout of all antennas	• Antenna type and size – Realizable gain, sidelobe level and cross-pol level
	• Deployment mechanism design/selection	• Deployment and release structures – On-orbit thermal distortion and scattering
	• Pointing error	• Antenna pointing loss and gain stability
On-orbit environment	• Temperature range and thermal distortion	• Thermal stable material (graphite material) and structure design – Reflector hydroscopic effect – On-orbit thermal distortion – Depolarization – RF loss • Thermal blanket and sunshield – RF loss
Manufacturing tolerance	• Surface accuracy	• Co-pol and cross-pol degradations
RF self-compatibility	• Mutual coupling	• Minimum physical separation and/or field of view (FOV) clearance – Sidelobe and/or cross-pol performance – Ring around interference
	• Passive inter-modulation (PIM)	• Passive intermodulation (PIM) control – Sufficient Rx/Tx port-to-port rejection in OMT/diplexer – Great attention to control design/manufacturing/testing process and facility clearance

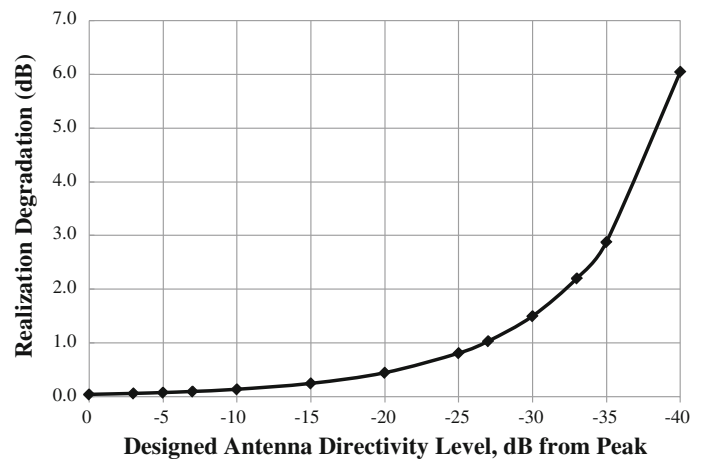
variation of the insertion loss and group delay with frequency in order to minimize the signal distortion. The channel signals emerging from the input multiplexer will be amplified by a nonlinear power amplifier, which introduces AM/PM and AM/AM distortion. Thus, any small variations of the signal amplitude and/or phase (group delay) introduced by the filters will result in distortion of both the amplitude and phase (group delay) of the signals at the output of the power amplifier. To minimize the variation, particularly the group delay variation, group delay equalization is usually required for IMUX filters, implemented either by using self-equalized filters or using external equalizers. Figure 14.45b depicts the insertion loss and group delay

flatness within the pass band of a 36 MHz IMUX filter with self-group delay equalizers.

3. Temperature stability. All the performance parameters (rejection, insertion loss, and group delay flatness variation with frequency) must be maintained over the environmental temperature change range to which the IMUX will be exposed in orbit. Since the fractional bandwidth of the individual IMUX filters is small (of the order of 1 % or less), any small drift or change of the center frequency of the filter with temperature can be a significant fraction of the filter's bandwidth. For example at C-band, a nominal 36 MHz bandwidth filter has a fractional bandwidth of 1 %. If the filter is constructed from aluminum cavity resonators with a coefficient of

Table 14.7 Typical Ku-band reflector antenna loss budget

Contributor		Deterministic (dB)	Random (dB)
<i>Feed assembly</i>			
1	Horn sunshield	0.02	0.006
2	Feed assembly insertion loss	0.22	0.060
3	Feed assembly mismatch	0.06	0.006
<i>Reflector</i>			
1	Reflector sunshield 2 way pass through	0.04	0.012
2	Reflector surface reflectivity	0.05	0.012
3	Reflector surface thermo-elastic distortion		0.250
4	Reflector surface hygroscopic effect		0.060
5	Reflector surface manufacturing errors		0.350
<i>Scattering</i>			
1	Combined scattering effects from spacecraft		0.120
<i>Other uncertainties</i>			
1	Modeling uncertainties		0.200
2	Measurement uncertainties		0.150
Total loss		0.39	0.52
Grand total loss		0.91	

Fig. 14.40 Typical realization degradation of low level co-pol and cross-pol performances of a Ku-band satellite antenna

thermal expansion of $20 \times 10^{-6}/^{\circ}\text{C}$, its frequency shift over a temperature range variation of 75°C would be about 6 MHz, or about 15 % of the filter's useful bandwidth. Thus, a much more temperature stable material is needed or temperature compensation of the frequency drift must be applied. Both of these techniques have been employed in the design and construction of IMUX filters. Invar, which has a thermal coefficient of expansion about $1 \times 10^{-6}/^{\circ}\text{C}$ has been used in the past, but it suffers from being heavy and difficult to machine. The current state-of-the-art IMUX filters use dielectric-loaded resonators (DR) for their realization [29]. The dielectric material used in these resonators is an extremely thermally stable ceramic, has very low loss, and a high dielectric constant (36–80). As a result, very small

size, thermally stable, high-quality filters are realized. Figure 14.46 shows a 3-channel Ku-band IMUX assembly. The individual filters used in this multiplexer are 10-pole self-equalized dielectric resonator filters.

14.5.5 Output Multiplexers

The functions of the output multiplexer (OMUX) are

1. To combine the amplified channelized signals from the output power amplifiers into one output port to feed the transmit antenna.
2. To minimize adjacent channel interference due to the spectrum regrowth as a result of the nonlinearity of the power amplifiers.

Fig. 14.41 A typical input filter assembly

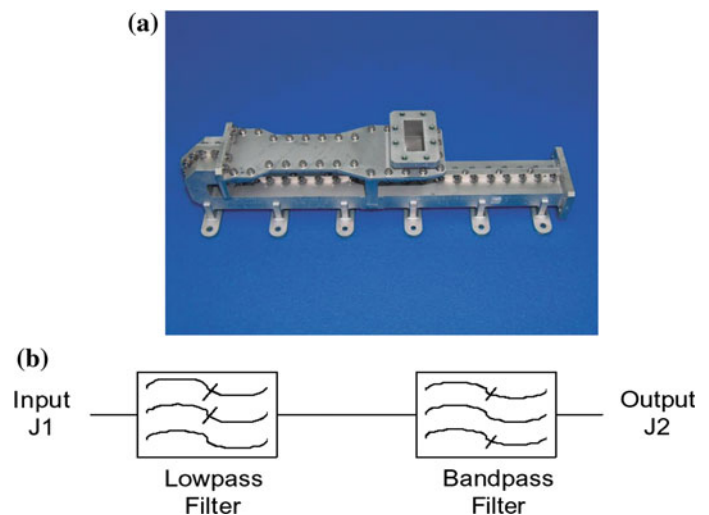
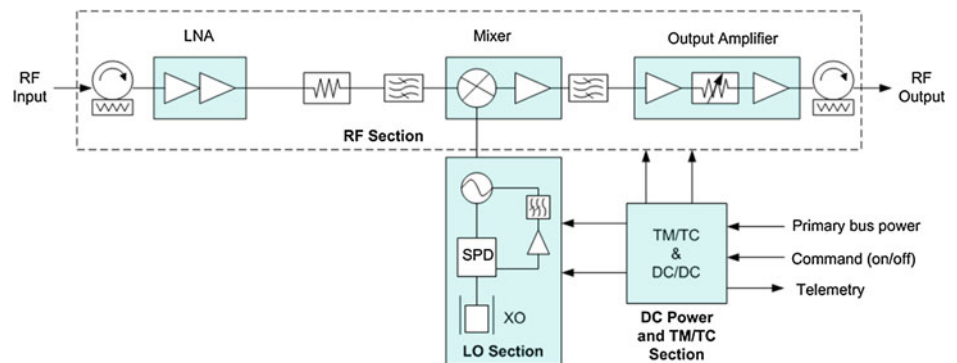


Fig. 14.42 Functional block diagram of a typical receiver



3. To reject the harmonics that are generated from the high power amplifiers.
4. To reject other spurious out-of-band signals (e.g. intermodulation products) generated by the high power amplifiers.

The important performance parameters of the OMUX are

1. Low loss. Since the OMUX filters carry the high power output from the power amplifiers the OMUX loss directly reduces the EIRP.
2. Power handling capability. For a typical 12-channel Ku-band OMUX connected to twelve 125 W TWTAs each filter must handle at least 125 W, and the OMUX must be able to carry more than 1.5 kW of RF power. The power dissipation due to the filter losses will cause a temperature rise in the OMUX. There must be sufficient cooling to limit the OMUX temperature rise. Usually the OMUX is mounted on heat pipes to facilitate this heat removal. Related to the power handling, the OMUX filters must not sustain any multipaction breakdown.
3. In-band performance. The pass bands for each of the channels must exhibit minimal frequency variation in order to minimize signal distortion.

4. Temperature stability. All the performance parameters (rejection, insertion loss, and group delay flatness variation with frequency) must be maintained over the environmental temperature change range to which the OMUX will be exposed in orbit. Most output multiplexer filters were manufactured using silver plated Invar alloy resonators that maintain a high degree of temperature stability and low insertion loss. At C-band, the state-of-the-art is to use dielectric resonators for output multiplexers to improve the temperature stability and reduce the size and mass of the OMUX. In the Ku-band, temperature compensation schemes using aluminum resonators have been used to minimize temperature drift and to reduce mass [30, 31]. DR filters have also been developed for Ku-band OMUX to further reduce size and mass.

Most OMUX designs use a rectangular waveguide manifold, with dual-mode circular waveguide resonator filters mounted on the manifold (E-plane or H-plane) [32], as shown in Fig. 14.47. The filters used are 4-pole to 6-pole quasi-elliptic function filters. Figure 14.48 shows a typical overall response of a 12-channel OMUX.

Fig. 14.43 LO phase noise spectral density versus frequency offset

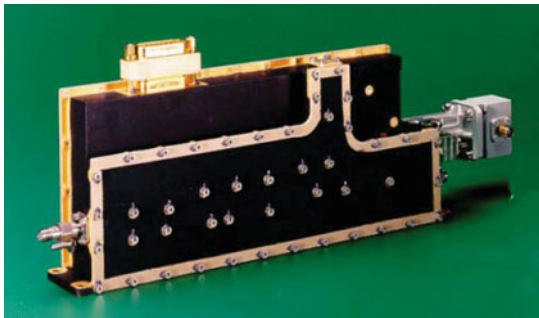
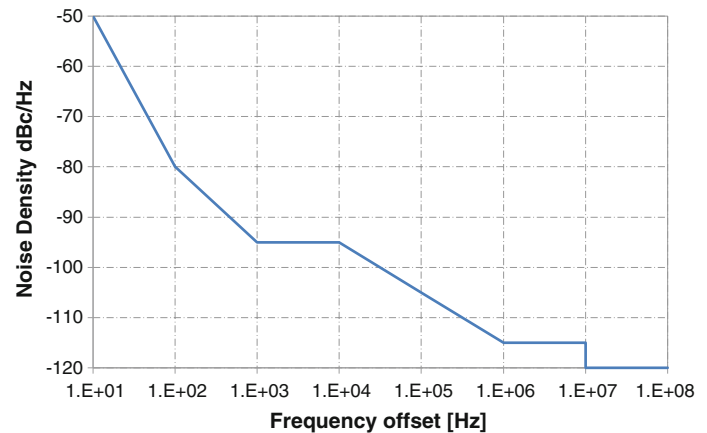


Fig. 14.44 A photograph of a flight C-band receiver. *Image TAS*

14.5.6 Traveling Wave Tube Amplifiers

The functions of the TWTA are to provide amplification of the RF signals with sufficient gain to achieve the required output power, with the least possible signal distortion (good linearity), highest possible DC to RF conversion efficiency, and high reliability over the design life of the satellite. The TWTA can consume most of the DC power resources of a satellite; hence, it is very important that they operate at the highest possible efficiency. This is achieved when the TWT is driven to saturation. To provide the necessary drive power to the TWT, a channel amplifier (CAMP) is required. In addition, a pre-distortion linearizer is usually used before the TWT in order to improve its linearity. The complete assembly of the channel amplifier, linearizer, TWT, and electronic power conditioner (EPC), as shown in Fig. 14.49, is often called an LCTWTA. The CAMP usually has two modes of operation, a fixed gain mode in which the gain of the CAMP is adjustable by ground command, and an automatic level control (ALC) mode in which the gain is automatically adjusted to maintain a constant output level (i.e. a constant drive level to the TWT) within a given range of the input signal level to the CAMP. The EPC regulates the DC power from the spacecraft power bus, and generates the necessary high voltages for the TWT.

As shown in Fig. 14.50, a TWT possess four major assemblies

1. An electron gun that produces a high density electron beam.
2. A microwave slow-wave circuit that supports a traveling wave of electromagnetic energy with which the electron beam can interact.
3. The collector that collects the spent electron beam emerging from the slow-wave circuit.
4. The TWT package that provides points for attachment to the using system, provides cooling for power dissipated within the TWT, and includes parts of the beam focusing structure.

Amplification in a TWT is attained by causing an electromagnetic RF wave to travel along a propagating structure in close proximity to an electron beam, as indicated in Fig. 14.50. At the left of the diagram is an electronic gun assembly. The cathode, when heated, emits a continuous stream of electrons. These electrons are drawn through an aperture in the anode and are then focused into a well-defined cylindrical beam by a magnetic field. The beam is thereby caused to travel inside the slow-wave circuit for the length of the tube. The electrons are finally collected and their kinetic energy is dissipated in the form of heat in the collector.

At the same time that the cylindrical electron beam is moving along the length of the tube axis, the RF signal to be amplified is fed into the slow-wave structure consisting of a coiled wire or a helix. The RF energy travels along the helical wire at the speed of light. However, because of the helical path, the energy progresses along the axial length of the tube at considerably lower axial velocity, determined primarily by the pitch and diameter of the helix.

The phase velocity of the RF wave (i.e. the speed at which the phase fronts of the energy appear to move along the length of the tube) is made slightly slower than the velocity of the electron beam. This near-synchronism results in a continuous interaction between the electron

Fig. 14.45 A typical IMUX frequency response for a 36 MHz channel

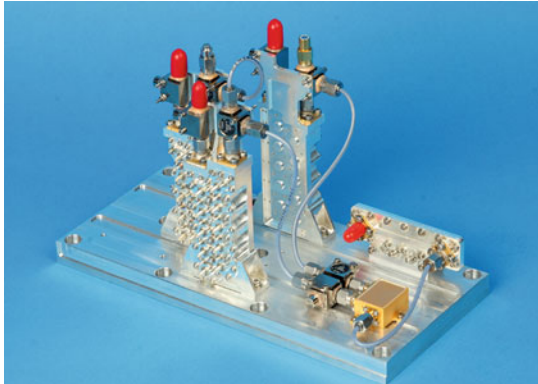
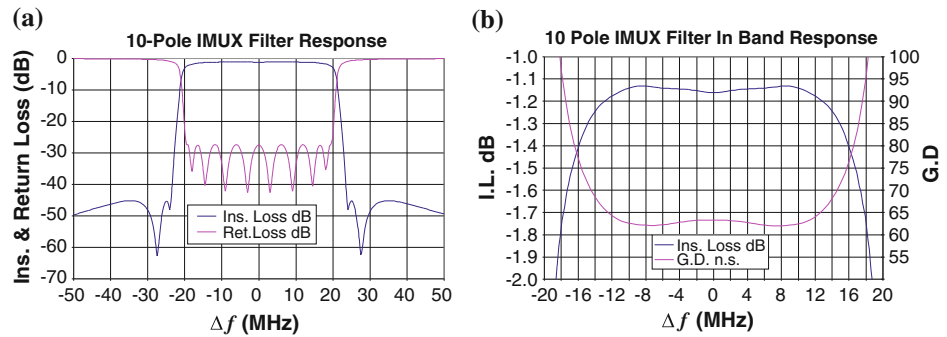


Fig. 14.46 A Ku-band IMUX assembly with DR filters. *Image TESAT*

beam and the RF signal. Some of the electrons in the beam are slowed by the RF field, while others are accelerated.

As the ‘velocity-modulated’ electrons move down through the helix they form bunches. These bunches, in turn, overtake and interact with the slower helical RF wave, surrendering kinetic energy to the wave on the helix. The result is a cumulative amplification of the RF signal. A typical state-of-the-art space qualified TWT has gain of 55 dB or more, the output RF power at saturation ranges from 10 W to over 200 W, and the DC to RF efficiency at saturation is close to 70 %. With an EPC efficiency around 95 % depending on the spacecraft bus voltage, the overall TWTA efficiency is greater than 65 %.

The dissipated heat from the TWTA is removed by conduction cooling, radiation cooling, or both. Figure 14.51 shows photographs of a radiation cooled TWTA and a conduction cooled TWTA.

Several parameters of the TWTA affect the communications system performance

1. Intermodulation distortion. When more than one carrier is introduced at the TWT input, a mixing or intermodulation (IM) process takes place. This results in

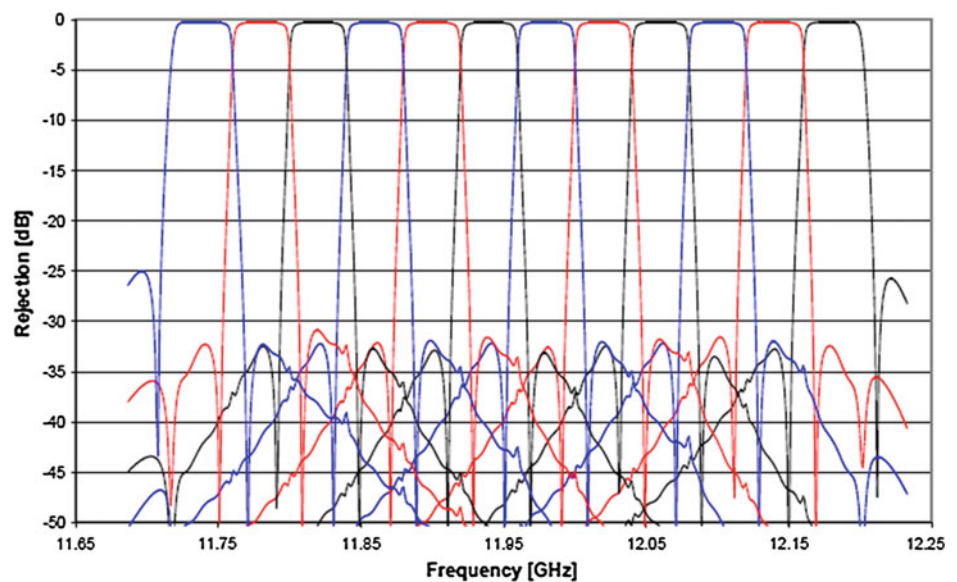
intermodulation products that are displaced from the carriers at multiples of the frequency difference. The power levels of these IM products are dependent on the relative power levels of the carriers and on the linearity of the TWT. In the case of two balanced carriers, Fig. 14.52 shows the variation of carrier and IM product power level with total drive power. The single carrier power curve is also plotted for comparison. The IM distortion is significantly reduced in the small-signal (linear) region of the RF drive range. For this reason communications TWTs are often operated backed off below their saturation levels.

2. Output power versus drive. The typical drive characteristic of a TWT is shown in Fig. 14.53. The minimum input signal level for useful operation (threshold) is determined by the bandwidth and noise figure of the tube. The linear dynamic range is that region between the threshold input level and the input at which there is a departure from small-signal or linear gain. The gain continues to decrease as the input level is increased, and is decreased by about 3–4 dB at the point of saturated output power. The DC to RF efficiency of a TWTA is maximum at saturation. As the drive level is backed off in order to improve the linearity, the efficiency decreases. To improve efficiency at higher output power levels while maintaining an acceptable linearity performance, a predistortion linearizer is added. The linearizer improves the TWTA performance by extending the linear drive region of the TWTA. Figure 14.54 shows the improved output power achieved by adding a linearizer. For the same C/IM value, the linearized TWTA has more output power (and DC to RF efficiency) than the TWTA alone.
3. AM/PM Conversion. Amplitude modulation/phase modulation (AM/PM conversion) is defined as the change in phase angle between the input and output signals as the input signal varies. This factor is measured statically and is expressed as degrees of phase shift per dB at a specified value of power output.

Fig. 14.47 Ku-band OMUX assembly integrated with redundancy switches. *Image* TESAT



Fig. 14.48 Typical frequency response of a 12-channel OMUX. *Image* TESAT



AM/PM conversion in a TWTA is caused by the reduction in beam velocity that occurs as the input signal level is increased and greater amounts of energy are taken from the beam and transferred to the input RF wave. At a level of 20 dB below the input required for saturation, AM/PM conversion is negligible. Beyond this point, AM/PM conversion increases sharply. A typical power output and relative phase shift response is shown in Fig. 14.53 for a TWTA. The phase shift is relatively insensitive to drive in the small-signal or linear portion of the RF output power characteristics. As the TWTA is driven towards saturation, the rate of phase change increases until saturation is approached and then decreases as the power saturates. The peak AM/PM generally occurs at a drive level 3–10 dB below the saturation drive, and is frequency dependent. The value of AM/PM conversion is less at the low frequency end of the tube's pass-band than at the high-frequency end.

Linearizers improve the AM/PM conversion characteristics of a TWTA.

14.6 Satellite Communications Systems Research and Development Trends

Due to significant market uncertainty over the lifetime of the telecommunications satellite and limited spectrum available for satellite communications, it is becoming increasingly important for satellite operators and/or service providers to (a) utilize the assigned spectrum more efficiently, and (b) build in-orbit reconfigurability and flexibility into their satellite fleets.

On the efficient use of the spectrum, most satellite communications providers have adopted digital communications. In addition, frequency reuse based on spatial and

Fig. 14.49 Block diagram of an LCTWTA

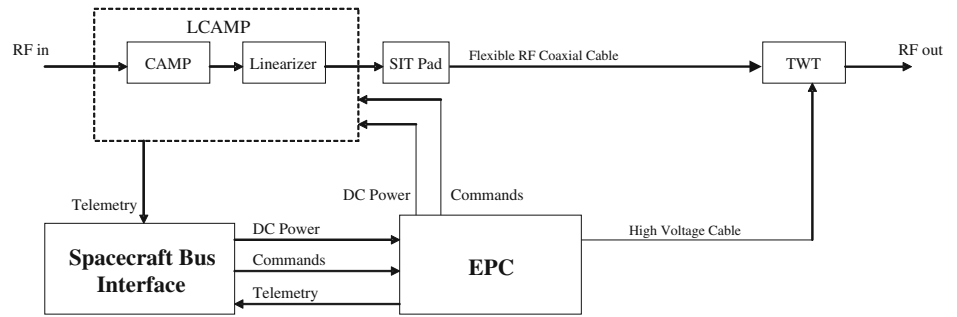


Fig. 14.50 TWT diagram and major assemblies

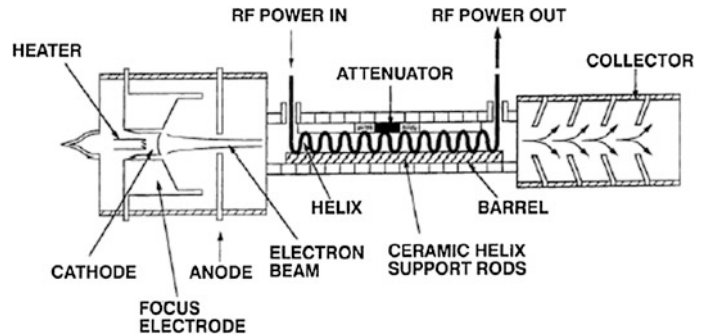
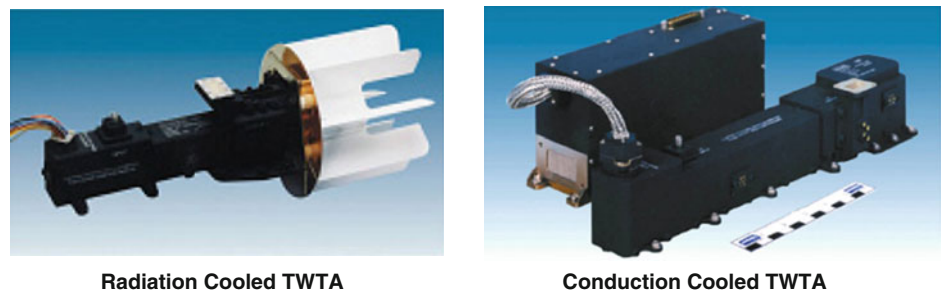


Fig. 14.51 Radiation cooled and conduction cooled TWTAs. Image L3 ETI



polarization diversities utilizing multiple regional and/or spot beams has become progressively more common for modern communications satellites. These techniques together have increased spectrum utilization efficiency many-fold compared with the analog communications and single beams used in the early days. Spectrum efficiency can be further improved by using on-board digital regenerative processors, which recover the transmitting signals on-board and therefore separate the uplink noise/interference influences from the downlink ones. Iridium and Spaceway satellites are two examples that utilize multiple spot beams and on-board digital processors.

Reconfigurability and flexibility are mainly focused into three areas, namely (a) frequency plan flexibility which includes channel bandwidth, frequency conversion and selectivity; (b) on-board power allocation/distribution flexibility which allows adjustment of the EIRP for given channels and/or beams based on the business needs within the capability of the satellite power subsystem; and (c) coverage

flexibility which principally concerns actions on the communications antennas. Reconfigurability and flexibility not only help satellite operators to manage their fleets more efficiently (including opening new orbital slots, covering new service areas, exploring new domains of applications in new frequency bands, and developing effective/inexpensive in-orbit and/or on ground back-up strategy) they also help satellite manufacturers to lower the nonrecurring cost and reduce manufacturing cycle for satellites since near identical payloads may be built and subsequently reconfigured either on-ground before launch or in-orbit by ground commands.

14.6.1 Frequency Plan and Channelization Flexibility

A major development in this area is the use of wideband agile frequency converters. This technology can be used together with on-board analog processors, digital

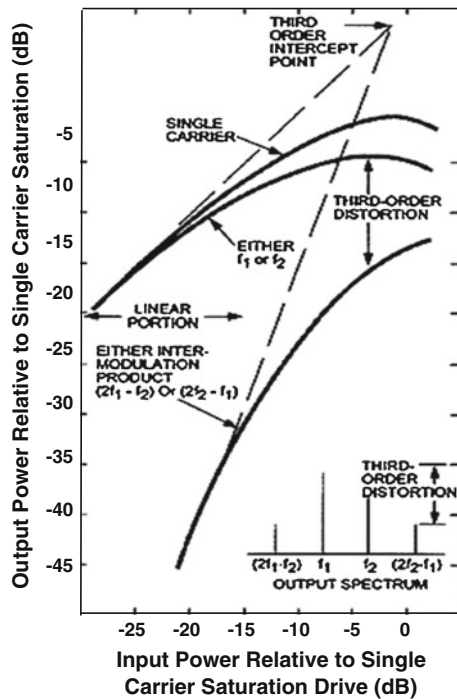


Fig. 14.52 Typical third-order IM data for a TWT

transparent processors (i.e. no demodulation/decoding), and digital regenerative processors to provide the required flexibility for channel bandwidth, frequency conversions, and channel/traffic routing/interconnectivity. As an application of this technology, a single-channel agile converter has been developed and applied to the Hylas-1 satellite. This equipment produces a flexible IMUX function and positions a fully variable bandwidth channel filter anywhere in the uplink or downlink frequency ranges by combining the agile converter with a fixed high performance low pass filter and a high pass filter, respectively. The IMUX channel filter response closely replicates that of conventional payloads and can be reconfigured to different bandwidths or center frequencies by ground commands [33]. In addition, an IMUX with tunable channel filters in the microwave band is also being developed. This technology, in conjunction with the wideband agile frequency converters, may provide a simple and more cost-effective solution for frequency plan and channelization flexibility.

Great in-orbit flexibility compatible with various frequency plans may be achieved without carrying multiple RF output multiplexers by employing

1. The technology for frequency plan and channelization described above.
2. Distributed amplification technology for combining multiple channels into a given antenna transmit port for downlink, such as multi-port power amplifiers.
3. Active array antenna technology.

14.6.2 Power Allocation/Distribution Flexibility

The key developments in this area include flexible TWTAs and gallium nitride solid-state power amplifiers (GaN SSPA).

14.6.2.1 Flexible TWTAs

Flexible TWTAs allow the saturation output power to be tunable within a limited range by controlling the anode voltage, while keeping constant high efficiency. Based on this concept, space qualified flexible TWTAs have been developed. Typically, the state of the art for the range over which TWTAs maintain nearly constant efficiencies is up to 3 dB. Compared to operating high power TWTAs at back-off, the flexible TWTAs significantly reduce DC power consumption at the same RF output power as illustrated in Fig. 14.55. This feature enables an operator either to operate more transponders within the same payload power envelope, or to boost the power in some transponders for customers requiring higher EIRP densities. In addition, from the satellite manufacturer's point of view, the flexible TWTAs also provide an opportunity for shortening the manufacturing schedule by inventorying common TWTAs for different RF power requirements of specific missions, since TWTAs represent long-lead items because they conventionally are tuned to specific frequencies and power in the design/manufacturing phase.

14.6.2.2 Gallium Nitride Solid-State Power Amplifiers

GaN SSPA technology, utilizing the properties of high power density and high junction temperature of GaN device, has demonstrated the capability of delivering much higher RF output power over a wide bandwidth with excellent DC power efficiency compared to the conventional LDMOS and GaAs technologies. The reported state-of-the-art results from various research/development institutes show that the saturated power in the C-band can be about 100 W with a power added efficiency (PAE) higher than 50 % and a linear efficiency in the Ku-band and Ka-band at least twice that of the corresponding GaAs SSPA. The advantage of GaN technology over GaAs will become even more significant at higher frequencies, such as in V-band. In addition, GaN SSPAs can be designed to provide near-constant efficiency and linearity with more than 3 dB RF output power variation by varying the drain voltage. This property, when used together with a multi-spot beam array antenna, provides the flexibility of power distribution among the beams equivalent to the one offered by multiplexed amplifiers (MPAs), but without the use of a Butler Matrix. This obviously will significantly reduce the system complexity and the output circuit loss.

Fig. 14.53 TWTA power output and phase shift as a function of RF input power

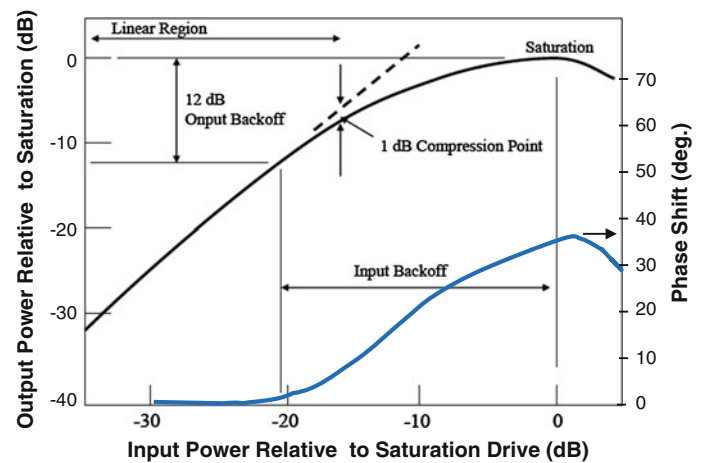
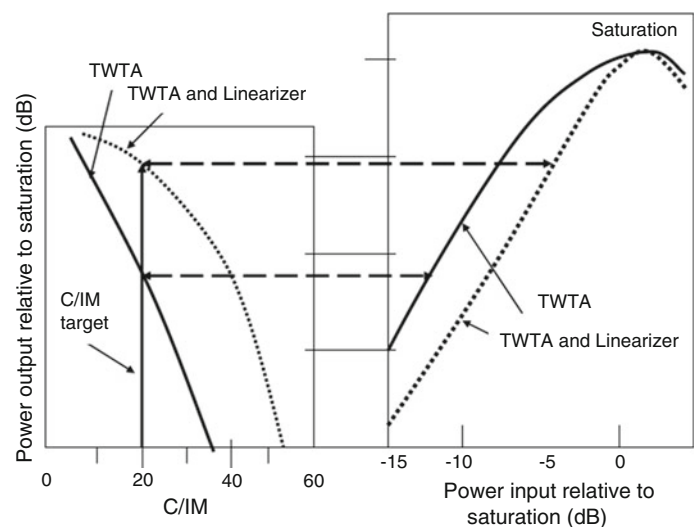


Fig. 14.54 Characteristics of a linearized TWTA



14.6.3 Coverage Flexibility

Many technologies for in-orbit reconfiguration of antenna coverage have been developed and/or studied. These technologies included those discussed below.

14.6.3.1 Feed Array Reflector Antennas

In the existing technology, the predefined beams are designed and built into the spacecraft antenna system with a different beam forming network corresponding to each beam. The in-orbit reconfiguration is achieved by on-board RF switches. This technology provides only a limited flexibility at the price of a more complicated and heavier spacecraft. Many telecommunications satellites have used this technology. Another means of in-orbit reconfiguration is to incorporate an electrically controlled low loss and high power phase shifter (for the transmit antenna) into each feed

element. The phase shifter is the critical component, and has been the focus of research and development.

14.6.3.2 Mechanically Reconfigurable Reflector Antennas

Mechanically reconfigurable reflector antennas in general can be realized either by mechanically switching in a pre-installed shaped main reflector/sub-reflector in orbit or by mechanically reconfiguring the surface shape of a reflector that was manufactured using a material with a memory or a deformable RF skin associated with small mechanical actuators.

14.6.3.3 Reflectarray Antennas

A reflectarray antenna combines some of the best features of microstrip array antenna technology and the traditional parabolic reflector antenna as shown in Fig. 14.56 [34, 35].

Fig. 14.55 DC power efficiency comparison between flexible TWTA and conventional TWTA operated at back-off. Image TESAT

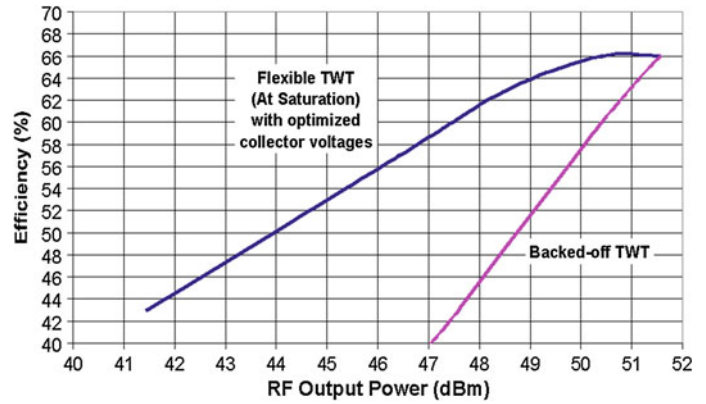


Fig. 14.56 Flat-plate microstrip reflectarray [34]

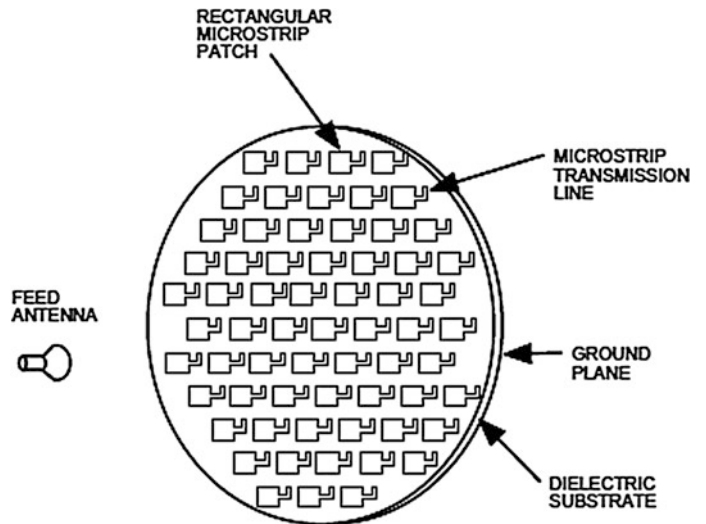
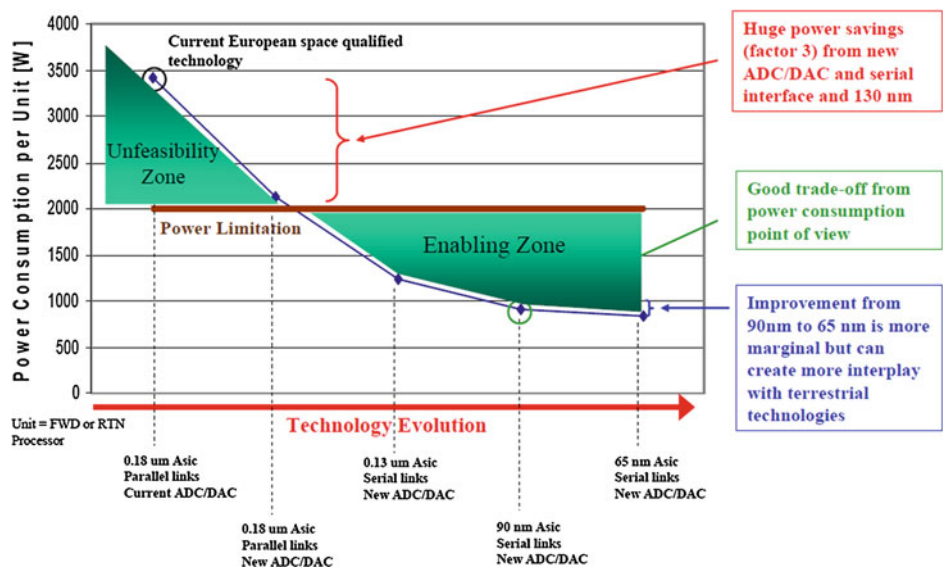


Fig. 14.57 Outlook of technology evolution for on-board wideband digital processors [38]



The reflector with printed array elements is usually flat or conforming to the required shape. The beam direction/shape can be controlled by a phase-delay line associated with each

element. With the progress in micro-electro-mechanical systems (MEMS) technology that enables the implementation of low loss micro switches in an integrated circuit, it is

feasible to control the phase delay of each element by switching in different phase-delay lines and therefore the antenna coverage can be reconfigured in orbit.

14.6.3.4 Active Array Antennas

An active array antenna with a digital beam forming (DBF) technology is capable of providing a complete coverage flexibility that includes beam shaping and beam steering [36]. However, for a communications satellite with a large number of beams, the on-board DBF will demand technologies near or beyond today's limit of the state of the art. As an attractive alternative allowing a high degree of satellite coverage flexibility while maintaining feasible satellite payload complexity, on-ground DBF technology was introduced with the NASA TDRSS geostationary satellites and then adopted by mobile satellite service systems such as ICO, MSV, and Terrestar [37]. The on-ground beam forming technology requires the transfer of radiating element signals to the ground and vice versa through a feeder link and the beam forming function is realized on the ground with all the flexibility offered by on-ground processing power. The available feeder link spectrum is one of the limiting factors for employing the on-ground beam forming technology to the wideband communications satellites.

14.6.3.5 Summary

All the discussed developments provide the short and medium term solutions to increase the flexibility and throughput of satellite communications. In the long term, the development and implementation of on-board wideband digital processors, coupled with on-board digital beam forming networks and high efficiency flexible HPAs, will make the high throughput and high flexibility satellite communications systems affordable. An outlook and evaluation of the key technology for on-board wideband digital processors is depicted in Fig. 14.57 [38].

References

- Noordung, H. "Das Problem der Befahrung des Weltraums - der Raketen-Motor", Berlin: Richard Carl Smidt & Co., 1928; published in 1928 with a date of 1929. First English translation: Hermann Noordung, Ernst Stuhlinger, J. D. Hunley, Jennifer Garland, "The Problem of Space Travel: The Rocket Motor", NASA SP-4026, NASA History Office, Washington D.C., 1995.
- Clarke, A.C., "Extraterrestrial Relays, Can Rocket Stations Give Worldwide Coverage?", *Wireless World*, Vol. 51, Oct 1945, pp. 305-308.
- Pierce, John R. (December 1990 (article)). "ECHO - America's First Communications Satellite". Reprinted from SMEC Vintage Electrics Volume 2 No. 1. Southwest Museum of Engineering, Communications and Computation.
- NTIA Web site: www.ntia.doc.gov
- Morgan, M. and Gordon, G. "Communications Satellite Handbook", Wiley 1989
- Morello, A. & Mignone, V., "DVB-S2: The Second Generation Standard for Satellite Broad-band Services", *Proc. IEEE*, Vol. 94, No. 1, p. 210 - 217, 2006
- Bishop, A. M., et.al. 'The INMARSAT4 Digital Processor and Next Generation Developments' *Proc 23rd AIAA International Communication Satellite Systems Conference*, Rome 2005
- Shannon, C. E. "Communication in the presence of noise." *Proc. IRE* 37, 1949: p. 10-21.
- ITU, "Handbook on Satellite Communications", 3rd edition, Wiley, New York, 2002
- ITU, Recommendation ITU-R P.372-10, Radio Noise, ITU, 2009
- ITU, Recommendation ITU-R P.676-8 Attenuation by atmospheric gases, ITU, 2009
- Crane, R. K., "Prediction of attenuation by rain," *IEEE. Trans. Commun.*, Vol. 28, No. 6, 1980, p. 1717-1733.
- Crane, R. K., "A two-component rain model for the prediction of attenuation statistics," *Radio Sci.*, Vol. 17, No. 6, 1982, p.1371-1387.
- ITU, Recommendation ITU-R P.618-10 Propagation data and prediction methods required for the design of Earth-space telecommunication systems, ITU, 2009
- Maral, G. and Bousquet, M., "Satellite Communications Systems", John Wiley & Sons, New York, 1982
- Schwartz, M. "Information Transmission, Modulation and Noise", Third Edition, McGraw-Hill, 1980
- Ziemer, R. E., and Peterson, R. L., "Introduction to digital communications", Macmillan Publishing Comp., 1992
- Rappaport, T. S., "Wireless communications principles and practice", Prentice Hall PTR, 1996
- Casini, E., De Gaudenzi, R. and Ginesi, A., "DVB-S2 modem algorithms design and performance over typical satellite channels," *Int. J. Satell. Commun. Network.*, vol. 22, no. 3, May-Jun. 2004.
- Assal, F., Mahle C. and Berman, A., "Network topology to enhance the reliability of communications satellites", *COMSAT Technical Review*, Vol. 6, No. 2, Fall 1976, pp. 322
- IEEE, "Antenna Standards Committee of the IEEE Antennas and Propagation Group, IEEE Standard Definitions of Terms for Antennas," *IEEE Std* 145-1973
- Gao, S., Clark, K., et., "Antennas for Modern Small Satellites," *IEEE Antennas and Propagation Magazine*, p. 40-56, Vol. 51, No.4, August 2009
- Mizuguchi, Y., Akagawa, M., and Yokoi, H., "Offset dual reflector antenna", *IEEE AP-S*, p. 2-5, Oct., 1976
- Rudge, A.W., Milne, K, Olver, A.D., Knight, P, "The Handbook of Antenna Design," Peter Peregrinus Ltd., London, 1982
- Ramanujam, P.; Law, P.H., "Shaped Reflector and Multi-Feed Paraboloid-a Comparison," *IEEE AP-S*, p1136-1139, Aug., 1999
- Kilic, O., and Zaghoul, A. I., "Antenna Aperture Size Reduction Using Subbeam Concept in Multiple Spot Beam Cellular Satellite Systems, Volume 44, No. 1, January 2009
- Rao, S. K., "Parametric Design and Analysis of Multiple-Beam Reflector Antennas for Satellite Communications," *IEEE Antennas and Propagation Magazine*, Vol. 45, No. 4, August 2003
- Chandler, C, Hoey, L., Smigla, T., Hixon, D., and Chan, R., "Ka-Band Communications Satellite Antenna Technology," 20th AIAA International Communications Satellite Systems Conference and Exhibit, Montreal, Canada; May 2002
- Kudisia, C., Cameron, R., and Tang, W-C, "Innovation in Microwave Filters and Multiplexing Networks for Communications Satellite Systems," *IEEE Trans. MTT*, Vol 40, No. 6, pp 1133-1149, June 1992

30. Cameron, R., Kudsia, C., and Mansour, R., "Microwave Filters for Communication Systems - Fundamentals, Design and Applications", Wiley 2007
31. Yu, M. "Power Handling and Temperature Compensation Design for Passive Microwave Devices", Proceeding of the 40th European Microwave Conference, Sept. 2010, pp 351-352.
32. Tong, R. and Smith, D. "A 12 channel contiguous band Multiplexer for satellite application'1984 IEEE MTT-s International" Microwave Symposium digest, pp 297-299
33. Thomas, G., Wheatley, N., Cobb, G., and Morris, I., "Agile equipment for an advanced Ku/Ka satellite", Flexible Payload Workshop, ESTEC, November 2008
34. Huang, J. "Microstrip Reflectarray," IEEE AP-S/URSI Symposium, Canada, p612-615, June 1991.
35. Georgiadis, A., Collado, A., and Perruisseau-Carrier, J., "Patents on Reconfigurable Reflectarray Antennas," Recent Patents on Electrical Engineering, Vol. 2, No. 1, p19-26, 2009
36. Litva, J.U., Lo, T. K.J., "Digital Beamforming in Wireless Communications," Artech House, 1996
37. Sichi, S., "Mobile Satellite Systems – A roadmap to Advanced Services and capabilities", 26th International Symposium on Space Technology and Science, Japan, June 2008.
38. Coromina, F., Mangenot, C., and Villette, E., "ESA Research activities in the field of Flexible Telecom Payloads," Flexible Payload Workshop, ESTEC, November 2008
39. Xiong, F., "Modem Technologies in Satellite Communications", IEEE Commun. Mag., p. 85-98, Aug. 1994
40. Sklar, B., "Digital Communications: Fundamentals and Applications", Prentice-Hall, 1988

Torbjörn Hult and Steve Parkes

The on-board data systems are responsible for collecting, processing, routing, storing and downlinking on-board generated data and for routing and storing uplinked data. Figure 15.1 shows a typical connection view of the data systems in a spacecraft. The dotted lines show optional connections and optional subsystems that may not always be present. The top row contains the subsystems that perform measurements and actions that either propel, change the orientation, heat/cool, or power the spacecraft. The bottom row contains the subsystems that perform communication with the ground facilities, i.e. telecommanding and telemetry. Although the Global Positioning System (GPS) receiver does not communicate with a ground station directly, it receives a ground maintained asset, time, via other spacecraft and its data can also be used to determine the satellite's position and attitude as discussed in Chap. 12.

The data systems communicate with the other subsystems using various communication links. These links may be point-to-point links, data buses or networks. The types and number of links used are determined by the spacecraft mission and the amount of data that is produced. Some communication between subsystems may use more than one type of link. The *power subsystem* may be controlled from the platform data system via a data bus but there can also be direct discrete pulse commands that control critical power distribution functions during emergency situations. The attitude and orbit control system (AOCS) sensors and actuators are often controlled via a data bus but very simple sensors and actuators are connected via individual links. The propulsion subsystem, on the other hand, often consists

of simple sensors and actuators and is thus almost exclusively controlled by discrete lines like valve open/close and valve position status lines.

The platform data system can be almost identical for many different missions. The only real difference is the interface to the AOCS subsystem where various orbits and satellite configurations require different sensors and actuators. The payload data system differs much more between missions. For missions having a single instrument the payload data system may sometimes not exist at all, or be integrated into the instrument. For telecom missions it is mainly a large data collector/command distributor managed by the platform data system. For advanced scientific or Earth observation missions it may include a complete data routing function, a large mass memory, and even a dedicated payload control computer.

15.1 Platform Data Systems

As mentioned, the platform data system is functionally the same for most spacecraft except launchers and human missions. Figure 15.2 shows the functional architecture. The functional architecture is the most generic way to describe the system, since the rapid evolution of hardware technology with ever more integrated circuitry results in different physical architectures. In the year 2000 a typical unit implementing a redundant data system, with discrete I/O for standard interfaces, AOCS actuators and sensors and for the propulsion valves and sensors, occupied 15.18 double Eurocard size (6U) boards. By 2011 this could be done in about 10 boards.

The main functions of the platform data system are

- Telecommand reception, decoding and handling including a direct ground capability of command pulse distribution.
- Telemetry Transfer Frame generation, coding and modulation with an optional essential telemetry sampler for

T. Hult (✉)
RUAG Space, Göteborg, Sweden
e-mail: torbjorn.hult@ruag.com

S. Parkes
Chair of Spacecraft Electronic Systems, School of Computing,
University of Dundee, Dundee, Scotland

Fig. 15.1 Platform and payload data systems in a spacecraft

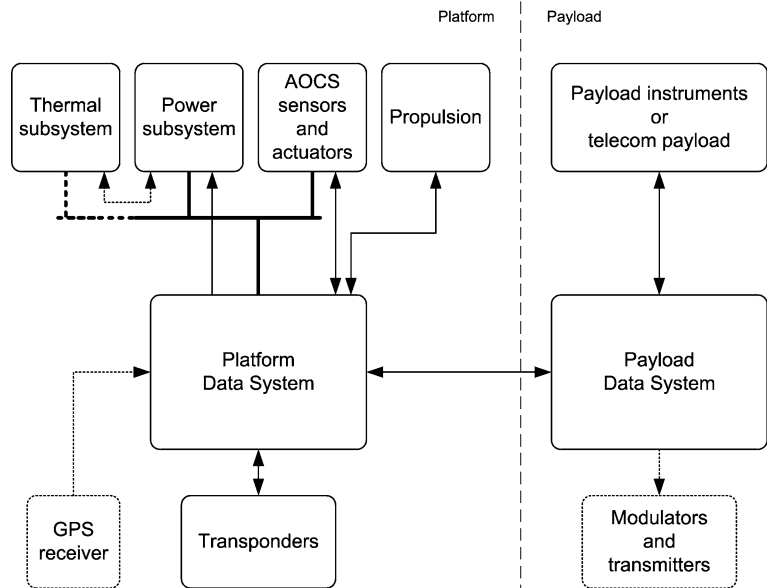
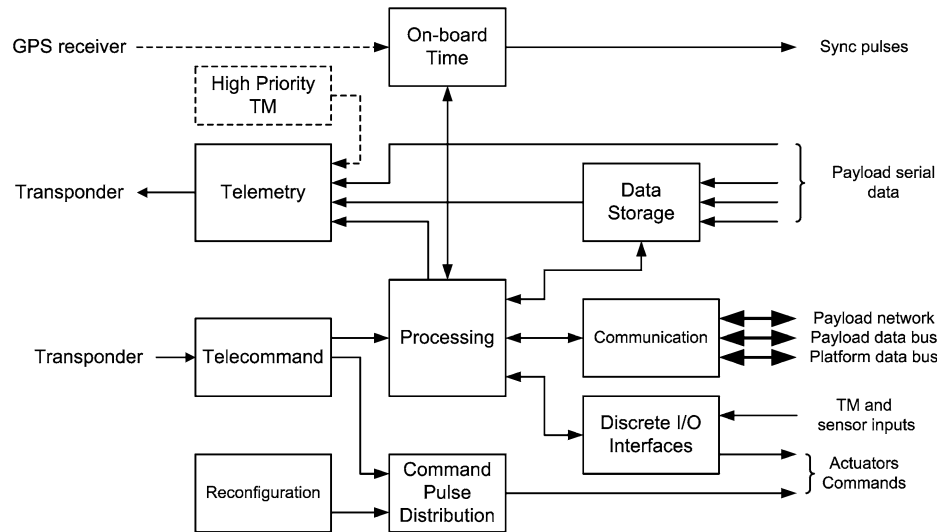


Fig. 15.2 Typical platform data system functional architecture



monitoring of various spacecraft equipment without involvement of the processing function.

- On-board time management, providing a stable time reference that can be synchronized to external events.
- Data storage capability for storage of telemetry and operational data.
- Processing capability with hardware drivers and operating system to store and execute application software.
- Communication with spacecraft platform and payload units.
- Discrete I/O interfaces for collection of on-board status and control data and for distribution of configuration and control commands.
- Fault detection, isolation and recovery (FDIR) in the form of a reconfiguration function that changes the current

configuration when errors are detected, and a safeguard memory that is used to store the current context of the processing function for later reuse in another configuration.

Each of these functions will now be described in more detail.

15.1.1 Telecommand

The spacecraft telecommand function is responsible for delivering command packets to on-board users. The command packets are generated on ground, embedded into various protocol layers in the ground station, uplinked via a radio link and received on-board by antennas and transponders. The functionality is quite well defined by international

standards. Since the 1980s the Consultative Committee for Space Data Systems (CCSDS), which is a worldwide organization, has prepared recommended standards for the space-to-ground communication protocols. Within the European Cooperation for Space Standards (ECSS) similar ground-space link communication standards have been written and these standards are compatible with the CCSDS standards, they merely narrow down the number of possible implementations. A good overview of all the communication links protocols is given in [1], while a more detailed overview of the telecommand protocol is described in [2] and formally specified in [3] and [4]. These standards include several figures of protocol layers and protocol data elements that are not repeated in this book, but the reader is recommended to have them available as supplementary information. The protocols are layered according to the principles established by the ISO open systems interconnection (OSI) model, with some minor differences. The main difference is that the OSI data link layer has been split into two sub-layers, the data link sub-layer and the channel coding and synchronization sub-layer, and equipped with two additional optional sub-layers for segmentation and authentication.

The channel coding and synchronization sub-layer [3] receives command link transmission units (CLTU) from one or more RF receivers depending on the specific configuration selected for the spacecraft. A CLTU begins with the start sequence, which is a pattern having good autocorrelation properties. The next part of the CLTU consists of data coded with a (63,56) Bose-Chaudhuri-Hocquenghem (BCH) code capable of correcting any single bit in the code word. A selection process selects which signal to use for further processing based on the quality of the start sequence and the quality of the code blocks. As the data is subject to pseudo-randomization process before being uplinked in order to ensure a sufficient bit transition density on the uplink, the decoded data from the code blocks are then subject to a de-randomization process before being sent to the next layer in the decoding process.

The data link protocol sub-layer [4] receives telecommand transfer frames from the channel coding and synchronization sub-layer. These frames contain a 5-byte header, which among various control information includes the address of the spacecraft and a virtual channel identifier (ID), a data field and a 2-byte cyclic redundancy check (CRC) field for further protection of the data. As the BCH decoding process may incorrectly correct code words that have more than two bit errors, there is a small chance that errors remain when the frame is processed. The frames are therefore checked for errors using the CRC. The spacecraft address is then checked and the data field routed either to an end user or to the next layer. If routed directly to an end user, the virtual channel ID can be used to determine the destination.

The optional segmentation sub-layer receives the data field from the data link protocol sub-layer. If implemented this sub-layer is always enabled. The data field is now called a telecommand segment and starts with a single-byte header that provides information on whether the segment is a standalone entity or a beginning, middle, or end segment of a larger data structure. The segment header also includes a multiplexer access point (MAP) ID, which is used to determine the destination of the telecommand segment.

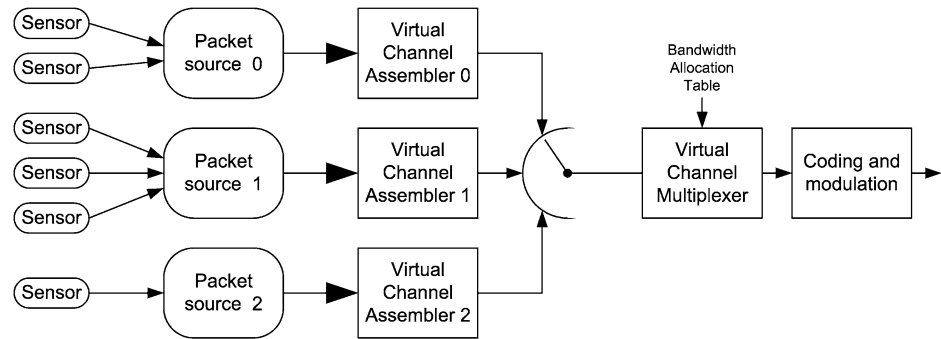
The optional authentication sub-layer receives the telecommand segments from the segmentation sub-layer and the data fields from the data link protocol sub-layer. It can often be enabled and disabled by telecommand. The first step in this sub-layer is to determine whether the commanding source is the correct one. This is done by comparing parts of an appended authentication tail by a calculated authentication code. This code is calculated from an on-board secret key also known by the ground station, from the telecommand segment data and a logical authentication channel (LAC) counter provided in the authentication tail. If the uplinked and calculated authentication codes are identical and if the uplinked LAC counter value is within an expected window, the telecommand segment (or the data field) is considered authentic and can be forwarded to the end user by the same mechanisms as described above. To prevent unauthorized access by someone recording and resending the same CLTU, the LAC counters are never reused with the same key. Keys are also replaced at regular intervals in case the key in use has been compromised.

The end user of a telecommand is in most cases the processing function. However, in some missions it is required to be able to perform essential telecommanding even if the processing function is not operating properly. In ESA missions, this function is called the command pulse distribution unit (CPDU) and allows sending a limited (typically 32–64) number of discrete pulse commands to switch essential spacecraft equipment. A CPDU receives a telecommand segment and extracts the embedded CCSDS space packet [5] that contains a list of one or more pulse commands to be generated.

The end user receives whatever data is embedded inside the telecommand segment. This typically is one or more CCSDS space packets, but it may also be an internet protocol (IP) datagram or other packet structure that can be uniquely identified by the first three or four bits that often determine the structure of the data received. The CCSDS standards define a specific encapsulation service [6] for those data structures that cannot be uniquely identified by the end user.

The various layers in the protocol stack allow implementation of a communications operation procedure (COP) between the ground station and the spacecraft [7]. A COP can be used to ensure that commands are always received

Fig. 15.3 Assembly of telemetry transfer frames



and accepted on-board a spacecraft in the same order as they are generated. This is done by using sequence counters in the telecommand transfer frame header, by only accepting transfer frames that arrive in sequence, and by down-linking in telemetry the next expected value of the sequence counter to detect whether a frame has been lost. For deep space missions a different protocol that does not guarantee the command order is often used because the communication delays do not allow receiving telemetry information about lost frames in due time.

When the telecommand function is made redundant, it is almost always operated in hot redundancy. The virtual channel ID is typically used by the ground operator to properly address the two telecommand functions. For implementations without the segmentation sub-layer, different sets of virtual channel IDs can be assigned to the two telecommand functions.

The implementations of the telecommand function are often governed by local end customer requirements. For instance, in ESA spacecraft that follow the ECSS standards, it is mandatory that the telecommand function as described in this section is completely implemented in hardware without involving any software. Other spacecraft implement the data link layer and upper layers completely in software, as is described later in this book. These spacecraft may have a device called a hardware command decoder that handles a limited set of command functions completely in hardware.

15.1.2 Telemetry

The spacecraft telemetry function is, like the telecommand function, quite well defined by international standards. The CCSDS and ECSS standards once again specify more details than can be described in this book. A detailed overview of the telemetry protocol is described in [8] and formally specified in [9] and [10]. As for the telecommand protocol, a layered structure with a data link sub-layer and a channel synchronization and coding sub-layer is used.

The data link sub-layer typically receives CCSDS space packets from a producer on-board the spacecraft. As for the telecommand protocol, any data structure is accepted since

the data link sub-layer does not make use of any data inside the data block to be transferred. The data producers are mainly the on-board computer application software, which generates real-time data, and the on-board mass memory, which stores packets during periods of ground station non-visibility. Each source of data is allocated one or more virtual channels on the downlink and for each virtual channel there is an assembly mechanism that sequentially packs the received data into fixed-length telemetry transfer frames as illustrated in Fig. 15.3. The structure of the frame is similar to that of the telecommand transfer frame. It starts with a 6-byte header followed by the user data stream. The telemetry transfer frame may end with a field called a command link control word (CLCW), which is used as reporting mechanism in a telecommand COP, followed by an optional 2-byte CRC for error detection purposes.

The assembly mechanisms sequentially number the generated frames using one numbering sequence for each virtual channel and temporarily stores them in a small buffer memory. The next step in the process is to select which frames to transmit. The most common principle used is a bandwidth allocation method that guarantees to each virtual channel a minimum available bandwidth. This bandwidth is automatically increased when other virtual channels do not fully use their minimum allocated bandwidth. Other principles may use priority for specific virtual channels. If there is no virtual channel ready to send data, the virtual channel multiplexer will start generating an idle transfer frame because the telemetry downlink relies on a continuous flow of equally sized telemetry transfer frames in order to keep the ground station reception process locked to the data stream.

The virtual channel multiplexer adds some information to the telemetry transfer frame header. The most important parts are the spacecraft ID and an overall frame count that is common for all frames irrespective of their virtual channel.

The channel coding and synchronization sub-layer receives the telemetry transfer frames from the data link sub-layer. To protect the data when being transferred on a noisy downlink the frames can be protected by an optional error-correcting code. Two main coding mechanisms are defined in the standards. The first code uses a non-binary

block code, a Reed-Solomon RS(255,223) code, that allows the correction of up to a maximum of 16 wrong bytes for each 255-byte block in the frame. The second code, a patented forward error correction code called Turbo code, give even better performance than the RS code but at the expense of a more complicated implementation.

After the coding process the frames, which are now converted to long code words, can be subject to an optional pseudo-randomization process in order to generate a sufficient bit transition density on the downlink. The next step is to prefix the frame with a 32-bit attached sync marker (ASM). Finally, if no coding or RS coding has been selected in the earlier stages, a final optional coding step called convolutional encoding can take place. This is a relatively powerful coding mechanism, especially in combination with the RS coding, that is very simple to implement on-board. It has been used for a very long time in space telemetry communication links.

The data structure now generated, starting with the ASM, is called a channel access data unit (CADU). These are continuously sent to the spacecraft downlink RF transmitter at a bit transmission rate that is determined by the available link bandwidth. The maximum available link bandwidth varies and depends on factors like the distance between the spacecraft and the ground antenna, the gains of the spacecraft and ground antennas, the elevation of the ground antenna, and the desired maximum error rate desired for the link. The application software can keep track of when it is possible to automatically change the bit rate, or the command to change this can come from the ground operator.

The receiving ground station continuously looks for the 32-bit ASM pattern. If two such patterns are found separated by exactly the number of bits that form the telemetry transfer frame (or the code word if coding is used) the ground station considers itself locked to the telemetry data stream and can start the decoding and user data extraction process needed before the on-board generated packets can reach their final destination on ground.

The telemetry function is, with very few exceptions, operated in cold redundancy and may also not be operated at all in order to save power during periods of no ground contact. The redundancy management is in most cases done by the ground operator, based on the quality of the received telemetry data. However, if users find that they cannot send their data to the telemetry function they may request an automatic on-board redundancy switchover. This switchover is then typically managed by an application software FDIR task.

The optional essential telemetry function is required to operate in the same way as the CPDU and hardware command decoder, i.e. to collect TM data and generate TM packets without involving the main application software.

The main reason is to avoid blind commanding in the event of software problems in order that the end result of all commands that use only hardware mechanisms can also be observed by hardware only.

15.1.3 Processing

The processing function is responsible for executing the application software of the spacecraft platform. The definition of what to include in the processing function varies but in addition to the pure computer hardware, it is quite common to include the hardware drivers and sometimes the operating system in the processing function, although this will also be discussed in [Chap. 16](#).

The processing function is the key element in modern data systems as more and more tasks are implemented in the application software and as the communication protocols normally handled by software, basically the ones above the data link layer, are becoming more and more complex.

The functionality provided by the processing is rather straightforward

- General processing capability, often from a general purpose processor and rarely from a specialized processor like a digital signal processor (DSP).
- Storage capability, both volatile and non-volatile, for application software code and data.
- A booting mechanism that loads the application software whenever the processing function is powered or reset. During the boot process it is often possible to run an optional self-test of the processing function.
- An application program interface (API) allowing the application software to communicate with the data system hardware.
- Scheduling mechanisms and event handling mechanisms allowing the application software to handle synchronous and asynchronous events.
- Timers to generate various events.
- Error detection mechanisms reporting hardware and software malfunctions.
- Error correction mechanisms that maintain correct memory content even in the event of transient errors.
- Development and debugging support allowing the application software developer to test and validate the software in its final environment.

In most spacecraft applications, the processing function is operated in cold redundancy. The main reasons for doing so are power consumption and reliability. An unpowered computer consumes almost no power and, as a rule of thumb, its failure rate is considered to be 10 % of the failure rate when powered. If the nominal processing function fails, the redundant one can take over within 5–20 s depending on whether a self-test is performed or not.

In some missions, the processing function is required to be operated in warm redundancy. This means that the redundant processing function is powered and executing some application software, and is ready to take over if the nominal processing function fails. The switch-over process is in this case much faster, of the order of 0.1–1 s. This mode can be required for interplanetary missions where for instance the planetary orbit insertion maneuvers are considered critical.

The processing functions can also be operated in hot redundancy, with both controlling the spacecraft by parallel reading of sensors and parallel commanding of actuators. This type of configuration is rarely used, as it doubles the probability of generating erroneous commands and complicates the selection of which command to use by an actuator.

15.1.4 On-Board Communications Links

The data system must communicate with other systems on-board the spacecraft. This can be done either by a centralized concept using dedicated communications links for each subsystem, or by using a bus or a network. The main principles for the three different concepts are shown in Figs. 15.4, 15.5 and 15.6.

The communications topologies have different properties and, depending on the needs, one may be more suitable for a specific application than the others. Table 15.1 shows a list of pros and cons. In most spacecraft to date, a bused concept has been selected since it provides a scalable and flexible interface for different spacecraft platforms. The bused concept also encourages reuse of standardized equipment between applications.

For future applications where higher overall data rates are expected, the bused topology will have a limitation. The data buses in use today have a transmission capability of less than 1 Mbps, and to increase the capacity more buses must be added. By going to a networked topology, transmission rates of tens or even hundreds of Mbps are possible.

To determine which link/bus/network to use a number of factors must be considered, primarily

- What types of data are to be transmitted? Single byte/word values or larger blocks of data?
- Which peak data rates and communication latencies are required?
- How many nodes are to be connected?
- Which inter-node communications paths are needed?
- Are there requirements for deterministic communications?
- Electrical aspects such as galvanic isolation, EMC and noise tolerance, power consumption and availability of space qualified interface circuits.

A more detailed description of some buses and networks in use today is found in Sect. 15.5.

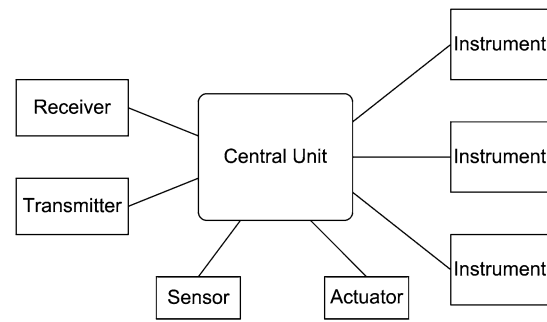


Fig. 15.4 Centralized communications concept

15.1.5 Sensor Data Collection and Actuator Commanding

The application software must have access to on-board sensors and actuators in order to perform its control tasks and to be able to monitor the status of the spacecraft. Most sensors and actuators can be classified as belonging to one of two types

- (A) ‘*Non-intelligent*’ simple devices that merely include the pure sensing/activating function. Examples of these devices are thermistors, contacts, motors, magnetorquers, and propulsion valves.
- (B) ‘*Intelligent*’ sensors that also include some kind of digital data processing.

Type A devices are typically connected to the data system via separate cables for each sensor/actuator and the conversion between analog and digital signals takes place inside the data system using a dedicated discrete I/O interface function. This function is called for instance remote terminal unit (RTU) or remote interface unit (RIU) when implemented in hardware. Type B devices include the analog/digital converter and are in many cases connected to the platform communication bus/network. It is expected that in the future there will be sensors that provide interfaces to so called ‘sensor buses’. These buses are simple links intended for low rate and short range communications between an RTU/RIU and a limited number of small sensors. Some corresponding commercial links are I²C, 1-Wire and LIN but so far there is no available standard for space applications, although work has been initiated with expected results within the ECSS system.

For type A devices there are numerous interfacing possibilities. An effort to standardize interfaces for relays, optocouplers, contacts, and thermistors has been done in for instance [11]. For other elements such as heaters, valves, motors, and magnetorquers there are no formal standards and each actuator manufacturer determines the interface characteristics. A *de facto* standard for actuators which need significant power has, however, emerged based on the most common primary power voltages available, 28 and 50 V.

Fig. 15.5 Bused communications concept

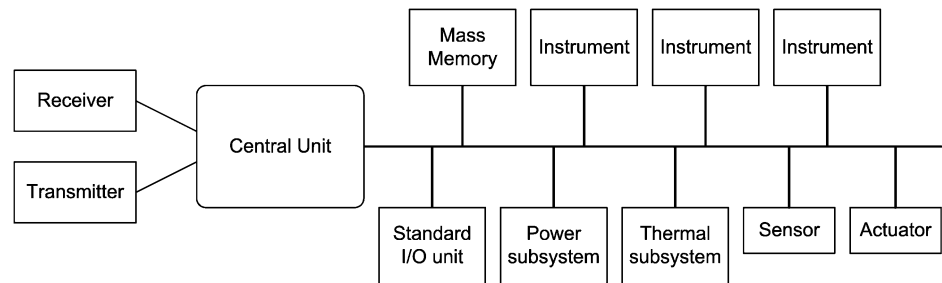
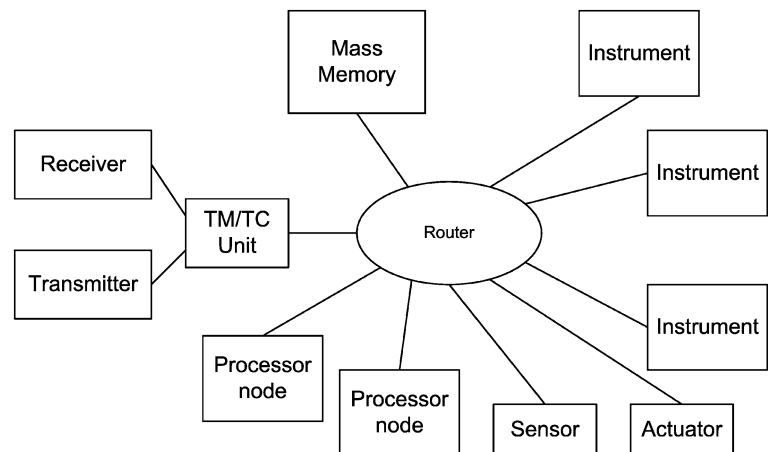


Fig. 15.6 Network communications concept



15.1.6 On-Board Time Keeping

The on-board time (OBT) function is responsible for maintaining the local spacecraft time. This can either be free-running or synchronized to a global atomic reference time like TAI; see Sect. 4.1.6.

A free-running OBT is typically used in launchers and other vehicles where the absolute value of time is not important. Even if the OBT is free-running, it is possible to determine the drift of the spacecraft clock relative to the UTC by for instance time-stamping specific telemetry frames, both when they are sent from the spacecraft and when they are received at the ground station. The time stamping on-board uses the OBT and this value is then transmitted as a data packet to the ground station within a few seconds after the event. By comparing the two time values and by knowing the distance from the ground station to the spacecraft, the OBT value can be determined with an accuracy of about 1 μ s. The method is described in more detail in [12].

Having determined the absolute error of the OBT, it is possible for the ground operator to synchronize the OBT by sending up telecommands that adjust the OBT value forwards or backwards. The adjustment can be done instantaneously, but it will have the effect that some OBT time values will be missing (in case of advancing the OBT) or will be duplicated (in case of rewinding the OBT). Having

missing or duplicated time values may upset some on-board functions that depend on preprogrammed time values for their execution. A better solution is to speed up or slow down the OBT clock. If the clock frequency is increased by say 1 % it will take 100 s to advance the clock 1 s and during this process there are no time values missing.

For LEO missions it is quite common to have a GPS receiver on-board. A GPS receiver generates a 1 Hz reference clock with an extremely good long-term stability, and can be used to keep the OBT synchronized with the TAI. If the GPS signal is lost, the OBT will continue to run on its free-running clock and can be re-synchronized when the GPS signal reappears.

For a free-running OBT there are in principle three classes of clock stability depending on the type of oscillator that is used

- For a standard crystal oscillator (XO) the OBT clock frequency will be within ± 50 to ± 100 parts per million (ppm) of its ideal value.
- For a temperature compensated crystal oscillator (TCXO) the OBT clock frequency will be within a few ppm of its ideal value.
- For an oven controlled crystal oscillator (OCXO) the OBT clock frequency will be within some tens of ppb (parts per billion) of its ideal value.

An XO has three main parameters. The first parameter is the initial setting error, which is typically a few ppm. The

Table 15.1 Comparison between different communications topologies

Architecture	Pros	Cons
Centralized	<ul style="list-style-type: none"> • Simple • Low cost in small systems • High data rates possible • Easy to integrate and test 	<ul style="list-style-type: none"> • Difficult to expand • Large systems become complex • Often many different types of interfaces • Difficult to monitor all traffic
Bused	<ul style="list-style-type: none"> • Modular and scalable to different requirements • Standardized interface to all units and subsystems Fairly easy to integrate and test <ul style="list-style-type: none"> • Easy to monitor all traffic 	<ul style="list-style-type: none"> • Complex for small systems unless the bus I/F is simple • It may be difficult to isolate a faulty device jamming the bus communication • Limited data rates
Networked	<ul style="list-style-type: none"> • Same as for the bused system (except the traffic monitoring) • High data rates possible • Distributed fault tolerance 	<ul style="list-style-type: none"> • Complex for small systems unless the network I/F is simple • Difficult to monitor all traffic • A faulty device could block other communication

second parameter is the temperature drift, which is a physical effect that depends on how the crystal is manufactured and is typically ± 20 to ± 30 ppm over the temperature range of -30 to $+70$ °C that is required for many spacecraft electronic units. The third parameter is the aging effect. This effect is most noticeable in the first weeks of operation, and decays with time. Typical values are <1 ppm/month for the first few months and <1 ppm/year after the first year.

The format of the OBT time code is also controlled by an international standard [13]. One commonly used format is the CCSDS unsegmented time code (CUC). It is a simple binary counter with the most significant part counting seconds and the least significant part counting fractions of a second. The maximum size defined is 32 bits of seconds and 24 bits of fractions. This gives a resolution of 60 ns and a wrap-around time of 136 years. The zero value of the OBT (also called the Epoch) can represent one of three commonly used moments in time

- The start of the mission. A common used name for this type of counting is mission elapsed time (MET).
- 1st January 1958, which is the date when the international atomic time (TAI) was synchronized with the universal time (UT).
- 1st January 2000; from the modified Julian date (MJD) introduced in Sect. 4.1.6.

15.1.7 Data Storage

The platform data system often has a data storage capability that is used to store platform telemetry data when there is no ground contact. This memory can also be used to store operational data like the mission timeline and on-board control procedures, and sometimes different application software images or configuration data, both for the data system processor and for processors in the payload.

The size of the platform data storage is typically in the range from a few Gbit up to about 32 Gbit. The latter limit is the addressing limit when using 32 bit addresses and also matches quite well what is needed when operating with the telemetry data rates in the order of a few Mbps normally provided by the platform data system. For a 1 Mbps downlink rate, a 32 Gbit memory will be downlinked in about 9 h. Thus, it is suitable for missions that do not generate data at high rates, such as interplanetary missions, where both the distance and the ground station coverage put a limit on how much data that can be downlinked.

There are three common ways to organize the data storage. The first way is to operate it like a classical hard disk with files and directories. Files can be dynamically created, deleted, opened, and closed, and they can grow and shrink dynamically. This mode is often combined with a file transfer protocol on top of the telecommand and telemetry protocols described earlier [14].

The second data storage method is to operate the data storage with a classical circular buffer concept. One or more circular buffers are created, and data can be written into a buffer and read from the same buffer in parallel. If ever the writing process is faster and the buffer is filled, there are two possibilities; either the writing process is halted until there is enough space available (a classical first-in first-out buffer, FIFO) or the oldest data is discarded and overwritten without ever being read. The method to use depends on the type of data stored and whether the data producer has additional buffering capability. A ground operations concept for circular buffers is described in [12] where the buffers are called packet stores.

The third data storage method is to use the data storage as a simple linear memory with no extra features.

The file storage concept can be used for both telemetry data and operational data. The circular buffer concept is only used for telemetry data, and the linear memory is mainly used for operational data.

15.2 Fault Detection, Isolation and Recovery, and Autonomy Support

One of the most important features of an autonomous vehicle is its capability of continued operation or safe operation after faults occur. In a spacecraft it is the task of the data system, in combination with the application software, to ensure that all spacecraft subsystems are sufficiently operational to meet the overall autonomy objectives. As the application software plays a major role in monitoring and controlling other subsystems, the main task of the data system is to keep the application software running. The mechanisms that perform this task are grouped into a common functionality called fault detection, isolation and recovery (FDIR).

15.2.1 General Dependability Terms

The definition of a fault tolerant system is generally not very exact, and can vary from application to application. There are however some commonly agreed key terms that are helpful in describing a fault tolerant system. A definition of the terms related to dependable systems can be found in [15] and is briefly summarized in this section.

- A system failure occurs when the delivered service of a system deviates from the specified service, where the specification is normally a combination of the required service and a description of the provided service (compare e.g. the supplementary information contained in a requirements specification and the corresponding user manual).
- The failure occurred because the system was erroneous, i.e. it contains an error. An error is typically a part of the system state that leads to the system failure. An error need not lead to a failure; it can be corrected by for instance some kind of error correcting code or another redundancy mechanism.
- The cause of an error is called a fault. A fault need not directly lead to an error; the error can be latent until it is activated. Faults are typically classified as physical faults and human-made faults, where the last class includes design faults and operating faults.

A dependable computer system is typically defined as a system that reliably delivers its services in the presence of faults and errors. Examples of dependability terminology are reliability, availability, maintainability, and safety, which essentially are explicit measures of dependability. Measures are not very useful if there are no means by which

to improve a system that has unacceptable measures. To achieve a dependable system, the set of means are

- *Fault-avoidance*—how to prevent, by design, the occurrence of a fault.
- *Fault-tolerance*—how to provide, by redundancy, the specified service in spite of faults occurring.
- *Fault-removal*—how to remove the presence of design faults.
- *Fault-forecasting*—how to estimate, by evaluation, the presence, creation and consequences of errors.

The remaining text in this section describes the mechanisms used to implement the fault-tolerance functionality.

15.2.2 Fault Detection and Isolation Mechanisms

As mentioned above the first step in the process of keeping the application software running is to detect faults that occur in the processing function. Since faults are detected as errors, the common term used is error detection. Some of the error detection mechanisms used by current implementations are

- Access to protected or unimplemented areas.
- Bus time out when accessing I/O devices.
- Error correcting codes in memory that detect correctable and uncorrectable error.
- Cyclic redundancy check and/or check summing of vital memory areas.
- Voting on multiple copies of vital memory data.
- Parity on address, data and control buses.
- Watchdog.
- Processor under-voltage detector.
- Built-in self-tests.

When errors are detected they can sometimes be mitigated immediately, as is the case of reading data from a memory that is protected by an error detection and correction (EDAC) device. An EDAC that detects a correctable error can still forward correct data to the processor. Note however that this does not mean that the fault in the memory is corrected. By reading the same address again the error will repeat.

If errors cannot be immediately corrected then some kind of alarm must be raised to signal that proper operation can no longer be provided. It is sometimes necessary to temporarily block the output from the processing function when an error occurs to prevent the error from producing unwanted output from the system.

15.2.3 Correction Mechanisms

The data system uses three main mechanisms to correct an error.

Memory scrubbing: The correctable errors in memory often result from a single-event upset (SEU). These are transient faults caused by a charged particle, a proton or a heavy ion, changing the value of one or more adjacent bits in a memory chip; see Sect. 3.3.2 for further information. Most EDACs have the capability to correct the effect of one SEU, but if nothing is done SEUs will accumulate in the memory and may result in uncorrectable errors. The solution is to continuously read the entire memory and, when a correctable error is detected, to rewrite the memory with correct data to remove the transient fault. This process is called memory scrubbing and can be done either by software or by a hardware mechanism.

Reconfiguration module (RM)—An RM is a function that listens to the various alarms generated by the processing function and then decides what action to take based on the nature of the alarm. A watchdog alarm is probably caused by malfunctioning application software and the first action in this case could be to simply reset the processing function. If this action is not successful, there is most likely a hardware fault somewhere and the next action would be to perform a switch-over to the redundant processing function. Other alarms, like under-voltage alarms or address parity error, are directly classified as generated by hardware and the first action is to directly perform a switch-over to the redundant processing function. The RM does not always rely only on alarms coming from the processing function. As errors cannot be detected with 100 % probability, there may be errors that result in no alarms but can cause incorrect behavior of the data system resulting, for instance, in loss of the attitude control function. A common method to detect even these possibilities is to have alarms originate from independent sources in various spacecraft subsystems. These alarms typically are

- Attitude anomaly
- Battery under-voltage
- Thermal limit reached.

In having these system alarms, it is considered that 100 % of the data system errors are detected, and the reliability modeling can be simplified since there are no single faults that can cause a system failure.

Safeguard memory (SGM)—When the RM has performed a processor reset or a processor switch-over, the application software that has started must be provided with some knowledge of the system status prior to the error. The mechanism used here is called safeguard memory, which is a memory that keeps its content during a processor reset or reconfiguration. During normal operation, the application

software regularly saves context data in the SGM and this data can then be retrieved by the software following a restart. Typical data that is saved are control loop state variables, actuator and sensor health status, and, of course, internal data system configuration parameters. The SGM is normally a volatile memory several hundred kilobytes in size that is permanently powered, but if the spacecraft can lose its primary power then some parts of the SGM can be implemented using non-volatile memory technology like electrically erasable programmable read-only memory (EEPROM) or magnetoresistive random-access memory (MRAM).

15.3 Payload Data Systems

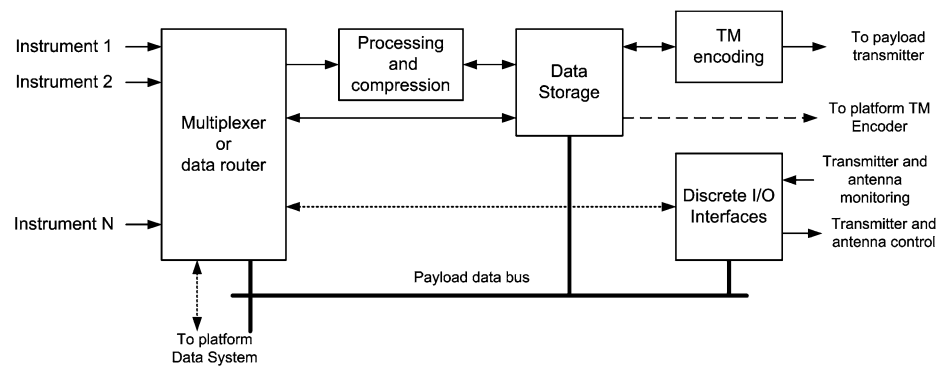
The architecture and functionality of the payload data system varies more than the platform data system. There are however some common elements and a typical payload data system for a spacecraft with more than one payload instrument is shown in Fig. 15.7.

The functions are

- **Instruments**—The payload instruments are the primary sources of data and the reason for the spacecraft being built. Typically a spacecraft carries several related instruments.
- **Data routing**—The data routing function forwards data from the payload instruments to one or more of the other common payload functions and comprises a multiplexer or router.
- **Data processing and compression**—An optional function that performs general signal processing, compression or data formatting tasks that have not been done in the instruments.
- **Data storage and telemetry encoding**—The data storage function stores data for downlinking when the spacecraft is visible from a ground station. If the spacecraft is in a geostationary orbit with 100 % downlink access, or data relay satellites are used, there may not be a need for data storage. The telemetry encoding function formats downlink data for missions that need dedicated high-bandwidth RF links to facilitate a high data output rate. If the payload data rate is low or medium, the data will instead be sent to the platform data system, where the telemetry encoding is done.
- **Discrete I/O interfaces**—A discrete I/O handling function that mainly manages the payload downlink RF system, which itself often consists of several units such as a modulator, high power amplifier, antenna pointing motor with associated drive electronics, and occasionally some RF switches.

The entire payload is connected to the platform data system via either a data bus or a high-speed link. If a data bus is used it is often connected to several functions inside

Fig. 15.7 Typical payload data system architecture



the payload data system. If a high-speed link is used, it is often connected to the router function, which allows full communication with all functions to which it is connected.

15.3.1 Instruments

The payload instruments on-board a spacecraft are the reason for the spacecraft being built. The instrument or set of instruments employed on a space mission will depend on the science drivers for that mission. There are three main types of mission

- Earth Observation missions, which orbit the Earth taking measurements of the atmosphere, oceans, land surface, topography etc.
- Science missions, which may orbit the Earth observing celestial objects free from interference from the Earth's atmosphere, or may travel to and possibly land on other objects in the solar system in order to explore them.
- Commercial missions, which include telecommunications and global positioning systems.

Instruments for Earth observation and science missions make measurements at various wavelengths of the electromagnetic spectrum and can be either passive, collecting radiation emitted by the observed body, or active, where the instrument emits an electromagnetic signal and collects the reflection of this signal from the observed body. Some example instruments will now be described.

15.3.1.1 Passive Optical Instruments

Probably the easiest passive instrument to understand is the optical imaging instrument or camera. The camera receives the light radiated or reflected from the surface being sensed, it focuses this light onto a focal plane to form an image, which is collected by an imaging sensor, for example a charge couple device (CCD). The imaging sensor usually contains a two-dimensional array of individual sensors, called pixels. When illuminated by incoming radiation, the pixel sensor will convert the photons to electrical charge. After a short period accumulating this charge, the whole image is read out. The

charge from the pixel is transferred to a special shift register that reads out the pixels one by one. As it leaves the imaging sensor, the charge for each pixel is converted to a voltage and then digitized to produce the required digital image.

The imaging sensor may contain a more or less square arrangement of pixels (e.g. $1,024 \times 1,024$ pixels) in which case the entire image is collected in one go, as is the case with a normal camera. Alternatively, the image may be long and thin (e.g. $17,000 \times 1$ pixel). This type of sensor is often used in Earth observation applications. In this case, the CCD array is organised with its long axis across the direction of motion of the spacecraft and it images a line on the ground across the track of the spacecraft. An image line is taken and transferred out of the CCD. In the meantime, the spacecraft moves forward, so that the CCD next views a line on the ground adjacent to the first line, and so on. The projection of the CCD on the ground (the area being imaged at one time) is like the head of a broom sweeping over the surface, so this is called a push-broom imaging sensor.

An imaging sensor can capture spectral information (color) using filters, a prism, or a diffraction grating to separate out the different spectral bands required. A separate CCD sensor may then be used for each spectral band. When a push-broom sensor is being used, the spectral information is typically spread across the CCD sensor in the along-track direction, e.g. using a prism, with the different spectral bands falling on different pixels of the sensor. In this case several pixels are required in the along-track direction, so the CCD may be manufactured as, for example, a $17,000 \times 32$ pixel sensor with the 32 pixels in the along-track direction giving 32 different spectral bands. An imaging instrument that records an image, sensing light in each pixel across its entire spectral range is called panchromatic. An instrument that separates the image into a small number of spectral bands is called multispectral, and one that records the image into many spectral bands is called hyperspectral.

As well as imaging visible light, an imaging sensor is also able to record images in the ultraviolet and infrared. Imaging in the thermal infrared is also possible, but requires the imaging sensor to be cooled.

The CCDs doing the imaging have to be placed on the focal plane of the optics, so that they receive a correctly focused image. The assembly comprising the CCDs and related read-out electronics is thus called the focal plane assembly. If a single CCD does not have a sufficient number of pixels to cover the area required then several CCDs will be butted together to form a larger CCD.

15.3.1.2 Passive Microwave Instruments

Molecules in the atmosphere excited by incoming energy (e.g. thermal energy or radiation from the Sun) will emit radiation when they drop back to their normal state. This radiation is emitted at a specific frequency or set of frequencies depending on the particular molecule, giving each type of molecule a unique spectral signature. If the spectrum of the signal received by a radiometer is plotted then it is possible to determine the atmospheric constituents from their spectral signatures. Wind speed can also be measured from the Doppler shift of the spectral signatures. A passive microwave radiometer comprises an antenna to collect radiation from the Earth's atmosphere, a front-end receiver that converts the microwaves into an electrical signal, and a back-end spectrometer that measures the spectrum of the received signal. The signal from the atmospheric components is very weak, especially compared to receiver noise. To increase the signal to noise ratio and make the spectral signatures of the atmospheric constituents visible, the received signal must be integrated for long periods.

15.3.1.3 Active Microwave Instruments

Active microwave instruments illuminate the target area with a microwave signal and then collect the signal reflected back from the target. A radar altimeter is an example of a relatively simple active sensor. It emits a pulse of radio frequency energy, which propagates towards the target body. This pulse hits the surface of the target body and some energy is reflected back towards the instrument. This reflected energy is collected by the instrument's antenna and the time for the pulse to propagate from the spacecraft down to the target body and back again is measured. Since the speed of signal propagation is known, the distance from the spacecraft to the target body can be determined. For a planetary lander this distance above the surface is used to deploy parachutes etc., at the right altitude. For an Earth observation mission the orbit of the spacecraft is well known, so the radar altimeter can be used to measure the mean height of the Earth's surface or the ocean's surface within the footprint of the radar. In addition to measuring height, a radar altimeter is able to measure average wave height over the ocean from the slope of the leading edge of the radar return pulse, and wind speed from its amplitude.

Another active radar instrument is the synthetic aperture radar (SAR), which is able to provide images of the Earth's

surface both day and night, and even when there is complete cloud cover. It has an antenna that looks to the side of the satellite track across the Earth, and combines the pulse compression technique used in radar altimeters with a similar signal processing technique that takes advantage of the along-track Doppler shift of the radar signal due to the motion of the satellite, to compress the radar return signal in order to produce a two dimensional image.

In addition to instruments that operate passively by receiving natural radiation from a surface, and those that actively sense the surface by sending out a signal to illuminate the surface, there are some instruments that use signals of opportunity: signal sources that illuminate the surface but which do not originate from the instrument. An example signal of opportunity is the signal produced by a global positioning system (GPS) satellite or by a TV broadcast satellite.

15.3.1.4 Instrument Data

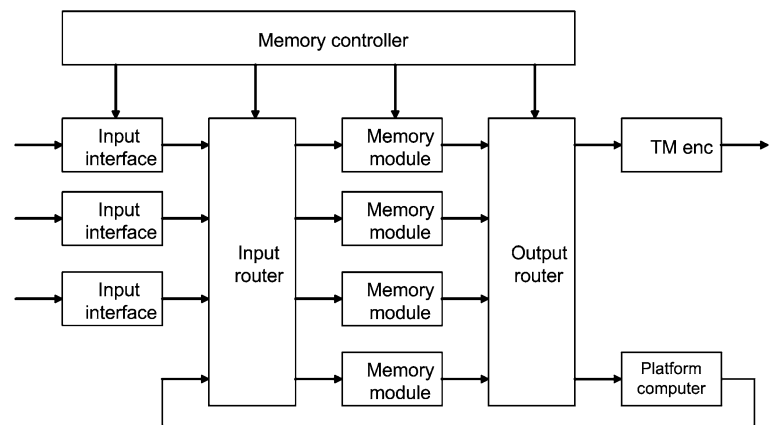
The data from an instrument clearly depends on the particular type of instrument. The data rate may vary from tens of kbits/s for some instruments like radar altimeters up to several Gbits/s for hyperspectral imaging and synthetic aperture radar instruments. On many spacecraft, a common data format will be applied to data from most, if not all instruments. Data from the instrument is packed into the common frame format before being passed to the data-handling system. This can simplify the handling of data in the mass memory unit. The data rate of the instruments will drive the architecture of the on-board data-handling system, with instruments requiring high data rates being connected directly to the mass memory unit and instruments with lower data rates being connected together by a bus or network and then attached to the mass memory unit.

15.3.2 Data Routing

The data routing function is responsible for passing data from the instruments to the data storage system, for passing configuration and control information to all of the instruments and payload data-handling units, and for gathering housekeeping information from the instruments. Instruments can have dedicated high-speed data links to the data storage system, or may share a bus or network resource. The discrete I/O interface could also be attached to the data routing function instead of being connected to a payload data bus.

The data routing function is typically implemented as a high-speed network with a central router or switch. A simple data multiplexer was rather common in former spacecraft but is becoming rarer owing to ongoing advances in communications link technology.

Fig. 15.8 Typical payload mass memory architecture



15.3.3 Data Storage and Telemetry Encoding

The payload data system has a data storage capability, which from a functional point of view is rather similar to the platform data storage function, i.e. it is used to store telemetry data when there is no ground contact. However, in some cases this memory is used to store other data. One example might be the payload memory in a planetary orbiter, which can be used to temporarily store data that is to be forwarded to rovers or to other spacecraft located on the surface.

The size of the payload data storage is in the range from several hundred Gbit up to a few Tbit. The larger memories are used for missions that produce large amounts of data at a high rate, like Earth observation missions with high-resolution optical instruments or synthetic aperture radar instruments. Since the large payload memories handle significantly higher data rates than the platform memories, their architecture is different. An example of a payload mass memory architecture is found in Fig. 15.8.

The data arrives from the instruments via dedicated input interfaces that perform initial data buffering and sometimes data compression. Due to the large size of the memory, there are several memory modules that are used in an M-out-of-N redundancy scheme in order to achieve the required reliability. An input routing mechanism forwards the input data to the active memory modules using either parallel buses or high-speed serial links. When one memory module becomes full, the router forwards the data to the next module in turn to receive data.

When data is sent from the mass memory it passes a routing mechanism that multiplexes data from the various modules and sends it to the downlink telemetry encoder. The payload telemetry encoder essentially works in the same way as the platform telemetry encoder but at significantly higher data rates. The output data rates are often defined by the capabilities of the RF modulators that typically operate at 150/155 or 300/310 Mbps downlink. A slightly different Reed-Solomon coding RS(255,239), is

often used to improve the coding overhead as there are only 16 extra code bytes instead of 32 bytes in the RS(255,223) code.

As discussed in Chap. 14, the telemetry encoder data is passed to the RF modulators, and then to the high power amplifier that generates the final signal sent to the downlink antenna.

To manage the entire mass memory operation it is common to have a dedicated memory controller. This handles the internal mass memory redundancy by powering on and off the various input modules and memory modules. It configures the memory module file system or packet stores, and sets up the output router and telemetry encoder with the proper parameters. It also commands the input router such that it knows which memory module to use when the current module becomes full. The memory controller is often based around a general purpose processor since it receives rather complex commands from the platform computer, either via the payload data bus or via a high-speed serial link.

Finally, there is normally an access path from the platform computer to the mass memory. This path is used to send housekeeping telemetry packets that are necessary to interpret the science data properly, like spacecraft attitude and position at regular points in time. For debugging purposes it is sometimes possible to read out the stored data to the platform computer.

For the payload data storage there are only two common ways to organize the data storage: file system or packet stores. The linear addressing sometimes used by platform mass memories is rarely used.

15.3.4 Data Processing/Compression

Processing of data to support the control or operation of an instrument is normally done within the instrument itself. There are some processing functions that are applicable to many instruments, or which do not involve any interaction

with the instrument control function. An example is data compression. These processing functions are normally implemented in a separate unit or board within a unit.

Imaging sensors, especially multispectral or hyperspectral ones, and synthetic aperture radar sensors provide very high data rates that cannot be supported directly by the telemetry downlink capacity. To cope with the large data rate, the data is stored in a mass memory unit and later downlinked to Earth at a slower data rate. This means that the high data rate instrument cannot collect data continuously, or the mass memory would rapidly fill up. Instead, a sensing schedule is devised so that the instrument can sense important areas of terrain, saving the data to the mass memory unit. It then has to wait until the data that has been recorded has been sent to ground, before it can collect any more data.

One way of increasing the effective downlink data rate and thereby reduce the time to send the information to ground, is to compress the data. For an imaging instrument this is normally done using JPEG or wavelet-based compression techniques. The degree of compression that can be achieved is typically of the order of 4:1 for Earth observation data. At this level of compression there is very little effect on the scientific information contained within each image, and it is possible to gather four times as much data. This is considered a good trade-off for many applications.

15.3.5 Discrete Input/Output Interfaces

The payload control function using discrete interfaces is rather similar to the platform sensor data collection and actuator commanding described in Sect. 15.1.5. For a scientific or Earth observation mission it mainly consists of monitoring and controlling the data downlink parts

- Modulator (temperature, voltage and single status bits)
- High power amplifier (temperature, voltage and single status bits)
- Antenna pointing motor with associated drive electronics
- RF switches (single status bits).

Apart from the antenna pointing motor electronics (APME), all these units are rather simple and the main interfaces are connected to Type A sensors and actuators like thermistors. The APME, on the other hand, can be rather complex depending on the antenna pointing requirements. In the simplest application, it just drives two stepper motors with the number of commanded steps per second in a commanded direction. This concept may induce unwanted vibration due to the rather fast acceleration and retardation of the motor. A more complex motor driver uses micro-stepping. When micro-stepping, a single motor step is divided into say 16 sub-steps, each of which forms one

point of a $\frac{1}{4}$ sine/cosine wave in order to give a smooth transition between two motor positions. This solution reduces the vibration problems and other problems that may occur when the stepper motor commanding frequency coincides with a mechanical resonance frequency. Moving large payload masses also affects the spacecraft dynamics and the AOCS system. Thus, the motor controller is often built as a separate unit and controlled from the platform computer directly via the payload data bus.

In telecom satellites the payload control function must control the many different units that make up the payload. The command/monitoring interfaces of these units are also rather simple, but the raw quantity will lead to hundreds and even thousands of inputs and outputs to the payload control function. The trend is therefore to decentralize the commanding and data acquisition as much as possible, and to run local data buses and data links between the payload controller and the units to control.

Scientific, optical, and radar instruments often include their own control and monitoring. In this way the instrument can be built independently from the payload data system and only interfaced by the high-speed data output, a control bus and a few simple interfaces like On/Off pulse commands, On/Off status, and maybe a temperature sensor.

15.4 Recent On-Board Data System Processing Technologies

The evolution of technology for data processing in terrestrial application has been very fast during past decades. According to Moore's law the number of transistors per chip doubles every few years (the specific number has been revised over the years) and the performance increases in a similar way. Although the maximum clock frequency of processors seems to have reached a limit, the performance is instead increased by putting more processors per chip. In terrestrial applications, quad core CPUs are now standard even in mid-range computers and mobile handsets.

The processors useful for space applications show a similar evolution, but with the difference that new processors appear at intervals of several years instead of a new processor every year as happens on the terrestrial market. However, the trend seems to be a lower rate of performance increase and there is a lag of 6–8 years between the commercial introduction of a technology in products and the introduction of that technology in space products. As the performance increase is slower, the lag is expected to increase.

The two following subsections describe the status of the main processing solutions for space in 2011.

15.4.1 SPARC Processors

LEON is an open source processor originally developed by the European Space Research and Technology Centre (ESTEC), part of the European Space Agency (ESA), and after that by Gaisler Research. The first widely used open version is LEON2, which has been extended with some fault-tolerance features to become LEON2-FT. The LEON2 processor is no longer supported, while LEON2-FT can be licensed from ESA to entities belonging to countries that are ESA members or procured as a packaged chip from Atmel (in two different versions).

The LEON3 processor is a further development made by Gaisler Research, now called Aeroflex Gaisler. LEON3 is also available in two versions; an open version under GNU GPL license and a commercial version called LEON3-FT. The most recent development is LEON4, which is only available in a commercial version.

The LEON processor family is based on the SPARC[®]-V8 architecture [16]. The main difference between LEON2 and LEON3 is a deeper pipeline and a single-edge clock design. The LEON3 support tools for both hardware and software design are also more advanced than the LEON2 tools. LEON4 has been further evolved using wider internal data buses and the possibility of having both level 1 (L1) and level 2 (L2) caches.

There are about ten different application specific integrated circuit (ASIC) implementations made by European companies like Aeroflex Gaisler, Astrium, and RUAG, using radiation-hardened or radiation-tolerant technology with the LEON2-FT or the LEON3-FT processors, with roughly an equal split between the two versions. Most of the implementations are available commercially from either the design house or an ASIC vendor. The designs use 0.25 μ or 0.18 μ process technology, and the performance ranges from 50 to 100 MHz clock frequency. LEON will be the dominating processor in European missions launched from 2014 onwards, after the Galileo and GMES programs that still uses the TSC695 SPARC[®]-V7 architecture processor. This processor has dominated the European market for space computers since the early 2000s.

15.4.2 PowerPC Processors

PowerPC was developed by IBM/Motorola/Apple. It has been used by Apple in desktop PCs but is today widespread in many embedded applications. PowerPC is an evolution of the IBM POWER architecture, and there are space processors developed from both architectures. BAE Systems has developed the RAD6000 processor, based on the IBM RSC chip that implements the POWER architecture. The

RAD6000 has been used in many missions, including seven missions to Mars. RAD6000 was manufactured in a 0.5 μ process and operates up to a maximum of 33 MHz.

BAE Systems has also developed the RAD750 processor, based on the IBM PowerPC 750, which belongs to the third generation of commercial PowerPC processors. The RAD750 is made in two different versions; in a 0.25 μ process operating up to 132 MHz and in a 0.15 μ process operating up to 200 MHz.

Honeywell has developed the RHPPC processor, based on the Motorola PowerPC 603e, which belongs to the second generation of commercial PowerPC processors. The RHPPC is manufactured in a 0.35 μ process and operates up to 80 MHz.

Maxwell has developed a board based on three IBM PowerPC 750FX processors operated in triple modular redundancy (TMR) such that an SEU will only interrupt the processing for about 1 ms. The board makes use of the fact that shrinking process dimensions of commercial circuitry automatically results in circuits tolerating rather large total radiation doses, thus needing no special processing for space applications. The design also benefits from the commercial IBM processor, processed in a 0.13 μ Silicon-on-Insulator (SOI) technology, being unusually insensitive to SEUs caused by protons and heavy ions. The CPUs are operated between 400 and 800 MHz.

15.4.3 MIPS: Microprocessor Without Interlocked Pipeline Stages Processors

The MIPS architecture was first introduced in the US with the RH32 and RH3000 processors and is now continued in Japan with the HR5000 CPU based on the MIPS IV architecture.

15.5 Recent On-Board Data System Communication Networks Technologies

The following subsections give a brief overview of some of the most common data links used in spacecraft applications. To get a deeper understanding and more detailed descriptions, it is recommended that the reader consult the relevant standards and handbooks. These open standards are often used as applicable documents in space projects and can be found either on the Internet or in company and institutional libraries and document data bases. It should be noted that both MIL-STD 1553B and SpaceWire are also discussed in [Chap. 16](#) from a software perspective, giving a slightly different view of these technologies.

15.5.1 MIL-STD-1553B

The bus defined by MIL-STD-1553B [17] is the most commonly used serial data bus in current spacecraft. It originated from the US Department of Defense in the 1970s and due to its widespread use in military aircraft and spacecraft for many years there is significant heritage and knowledge in industry. Even if it is more than 35 years old, the 1553 data bus still has sufficient data transfer capability for the control functions implemented in a spacecraft platform like attitude and orbit control, thermal control, battery management, and control of some payloads. The main advantages of the 1553 data bus are

- A single pair of wires is used, making it possible to use very small and simple connectors.
- Robust against noise and ground potential variations since it uses transformer coupled interfaces with high voltage levels and input signal filtering.
- Can be routed up to 30 meters, which covers the distances seen in even large spacecraft. When used in launchers there may be a need for bus repeaters to cover both the large distances and the separation between launcher stages.
- Sufficient availability of components that tolerate the space environment and that fulfill the quality requirements for space applications.
- Inherent redundancy features like dual bus lanes and rules for how to use the redundancy.

There are however a number of drawbacks

- The peak power consumption when transmitting on the bus is quite high. About 3 W of secondary power is needed and, as the power consumption while not transmitting is very low when using modern components, the variation in power consumption often results in electromagnetic compatibility (EMC) problems related to low frequency conducted emission on the primary power bus.
- The harness is rather bulky because it includes embedded transformers.
- The interface circuitry is complex and occupies more printed circuit board space than required by other buses.
- The interface circuitry implementation is also complex. There is an associated handbook [18] consisting of several hundred pages where one major topic is the electrical interface and all aspects to consider in its implementation.

Despite these drawbacks, it is expected that the 1553 data bus will continue in use in several future spacecraft.

15.5.1.1 Low Level Protocol Summary

The 1553 data bus connects two types of nodes, one bus controller (BC) and a number of remote terminals (RT). The bus uses a master–slave protocol where the BC is the

initiator of all transfers. Three types of transfers are possible: BC to RT, RT to BC, and RT to RT. The latter is rarely used and is not discussed further in this book.

All transactions, also called messages, start with the BC issuing a command word that determines the RT to be involved in the transaction, an RT sub-address indicating one of 30 separate register banks inside the RT, the number of data words to be transferred (from 1 to 32), and whether it is a BC → RT transfer or a RT → BC transfer. The RT is determined by a 5 bit RT address field where address 31 indicates a broadcast transfer, i.e. all RTs connected to the bus may receive data. It is not possible to transmit data from the RTs using broadcast addressing.

In the case of a BC → RT transfer, the BC starts to transmit the data words immediately after the command word. When all of the data words have been sent the bus is left idle for some time until the addressed RT responds with a status word that gives information on whether the transfer was successful or not. The RT must start its response within 12 μs from the end of the last data word.

In the case of an RT → BC transfer, the BC stops transmitting after the command word and the addressed RT responds with a status word that gives information on whether the transfer can be carried out or not, and then immediately transmits the requested number of data words.

In addition to the data transfers there are control commands, called mode commands. These are recognized by the RT sub-address being 0 or 31 and they may also result in a single data word being transmitted, either from the BC to the RT or from the RT to the BC.

All words are transmitted at a rate of 1 Mbps. A word lasts for 20 μs and consists of the following parts in order of appearance

- A unique pattern lasting 3 μs indicating whether it is a command word or a data word. The pattern is the same for command words and status words.
- 16 bits of command info, data, or status info.
- One parity bit generated from the 16 data bits. Odd parity is used.

Since the data bus is transformer coupled, the electrical signal level must be DC-balanced in order to avoid transformer saturation and overshoot problems when the bus transfers are completed. The signal modulation used is called Manchester coding or split phase level (SPL) coding and is a special case of binary phase shift keying (BPSK) with one subcarrier period per data bit.

15.5.1.2 Higher Level Protocols

The basic 1553 standard only defines how to transfer individual messages of up to 64 bytes. With such a protocol it is possible to control simple devices that only need or produce a few bytes of data. If larger data structures like CCSDS

space packets or files must be transferred then a protocol must be implemented on top of the basic 1553 message protocol. Traditionally such protocols have been invented by equipment suppliers and prime contractors. This has made it difficult to reuse software (and sometimes hardware) between various spacecraft.

In 2008 there was an international standard released within the ECSS system [19] specifying the communications layer above the data link layer. In the ISO OSI model this layer is called the network layer and [19] covers parts of this layer. The ECSS standard defines the following protocol services that use the basic 1553 messages as underlying protocol

- Time synchronization
- Time distribution
- Communication synchronization
- Data block transfer
- Terminal management.

By 2012, the ECSS standard was used in almost all new European missions having 1553 data buses.

15.5.1.3 Communication Scheduling

The 1553 data bus is a serial communications bus with a rather low speed. A full size message carrying 64 bytes occupies the bus for about 700 μ s and it cannot be interrupted. Thus, it is not possible for software to perform urgent random data accesses over the bus in order to access sensors and actuators. Instead, the bus traffic must be scheduled to meet the performance requirements in terms of bandwidth and latency. The latency is basically the time between the transfer request being made by the software and the transfer being completed on the data bus (Fig. 15.9).

1553 data bus engineers have classically scheduled the bus communication using so called major and minor frames. The duration of major frames is typically 1 s and is adapted to control functions that have the same periodicity, like time management, thermal control, and battery management. Some applications have also introduced ‘super frames’ with a period of, for example, 10 s to run the thermal control. The minor frame often coincides with the AOCS control loop, and is in the order of 100 ms.

With a minor frame having the same length as the AOCS control cycle, there must be some interaction between the software and the 1553 bus controller hardware during the frame, for instance to read the acquired sensor values or to write the generated actuator control commands. Most 1553 controllers have advanced list management mechanisms that facilitate transmitting multiple messages in a single software operation. Quite often these lists may even be triggered without software interaction, by hardware events generated for instance by the on-board time function. To avoid frequent interaction between hardware and software, [19] has introduced a third element, the communication

frame. The rationale for this frame is that the interaction between the hardware and the software only occurs at the frame boundaries, thereby simplifying both the software and the communications scheduling. This principle is shown by an example in Fig. 15.10.

In the example shown in Fig. 15.10, there are four communication frames per minor frame. A typical AOCS control loop would acquire the sensor data via the data bus during frame 0, perform the control algorithm processing during frame 1 and frame 2 and send out the control commands on the data bus during frame 3. During frame 1 and frame 2 the data bus transfers typically contain data related to other control functions like thermal control and monitoring. The resulting latency is a maximum of one communication frame, which is 25 ms in the example. If shorter latencies are needed a shorter communication frame should be selected.

The example shown in Fig. 15.10 also clearly shows how the data bus communication can cooperate with the software in a deterministic time-triggered mode instead of the earlier event driven modes that make the software less predictable and more error prone.

15.5.2 CAN: Controller Area Network

A controller area network (CAN) is a bus standard devised for automobile applications that allows various devices on the bus to send data to one another [20, 21]. It is a serial bus that typically runs over a 2-wire differential interface and provides a multi-master, message-based communications protocol. CAN has been used in many small satellites for payload data-handling and control applications, where the data rate requirements are low.

A CAN has a number of features that make it attractive for space applications

- It can send data at up to 1 Mbits/s over distances of 40 m.
- Any node can send a message to be received by one or more other nodes.
- Data delivered and accepted by multiple nodes is guaranteed to be the same, consistent data.
- Messages are delivered according to a fixed priority scheme.
- The latency of message transmission is fixed, providing time-synchronization of nodes receiving a message.
- Error are detected and signaled to other nodes.
- Messages that are corrupted are retransmitted as soon as the bus becomes idle.

A CAN is not galvanically isolated, although opto-couplers can be used to provide galvanic isolation.

The CAN protocols are divided into three layers

- CAN object layer, which determines when messages are to be transmitted and what received messages are to be

Fig. 15.9 1553 data bus messages

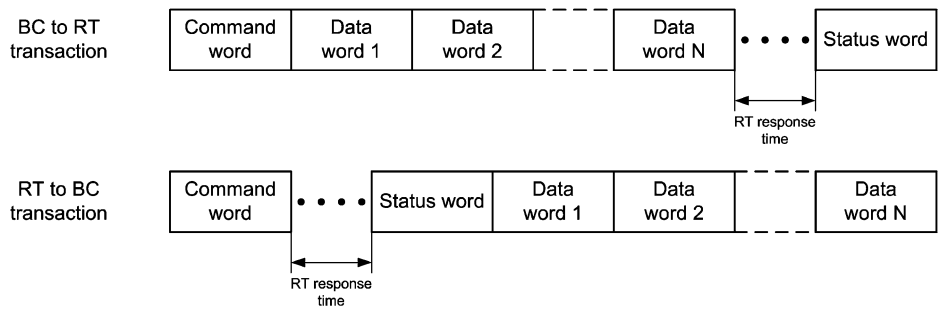
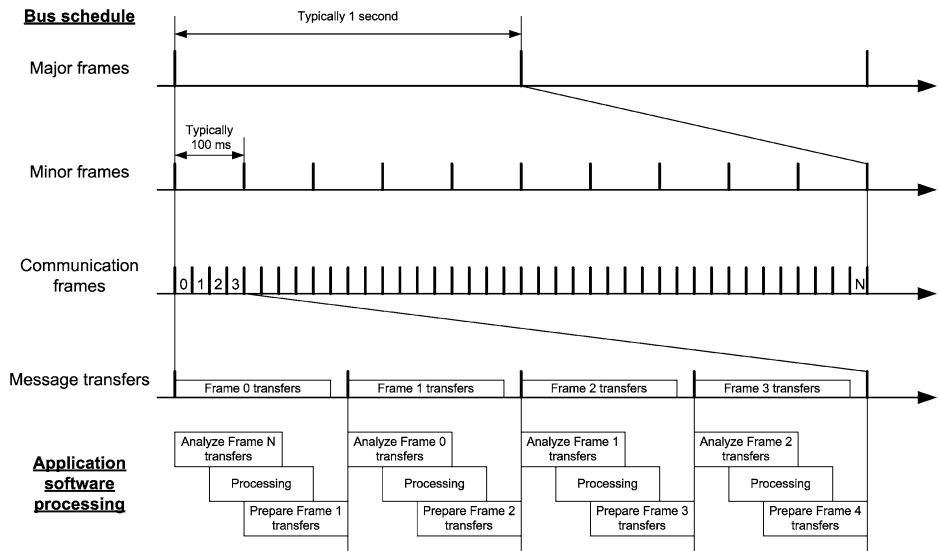


Fig. 15.10 1553 data bus transfer scheduling



passed up to the application. It provides the message handling and status interfaces to the application.

- CAN transfer layer, which is responsible for transferring messages between nodes. It provides bit timing, synchronization, message framing, arbitration, error checking, error signaling and fault confinement functions.
- CAN physical layer, which defines the electrical properties of the medium to permit the transfer of bits between nodes. Various forms of physical layer can be used.

Together the transfer layer and the object layer provide the services and functions of the ISO/OSI link layer.

There are four different frame types used in a CAN

- Data frames, which carry data from the transmitting node to the other nodes on the bus.
- Remote request frames, which are sent by one node to request one of the other nodes on the bus to transmit some specific type of information, this being subsequently transmitted in a data frame.
- Error frames, which are transmitted to indicate that an error has been detected.
- Overload frames, which are used to provide an additional delay between one data or remote frame and the next.

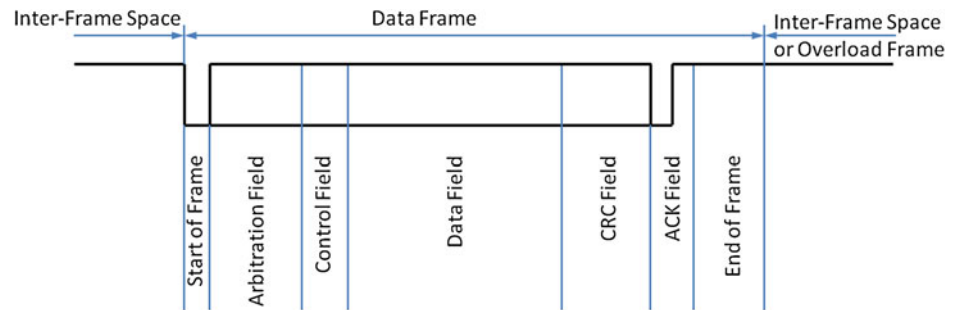
15.5.2.1 Data Frames

The data frame format is illustrated in Fig. 15.11. The start of frame is 0, which follows a series of 1's from the inter-frame space. A node on the bus is only permitted to start transmitting a frame when the bus is idle (1). The start of frame is used to provide initial bit-synchronization and frame synchronization.

The arbitration field contains an 11-bit field which identifies the type of information being broadcast, and where it comes from. No two nodes can transmit the same value in the arbitration field. In addition, the arbitration field contains a remote transmission request bit, which is 0 in a data frame, and 1 in a remote request frame. The arbitration field can also contain an extended identifier, which is 29-bits in total.

The control field contains an identifier extension bit that indicates whether the frame is a base frame (set to 0) or an extended frame (set to 1), a reserved bit (set to 0), and a 4-bit data length that indicates how many data bytes are contained in the data field (0–8 bytes).

The CRC field contains a 15-bit cyclic redundancy check sequence together with a 1-bit CRC delimiter, which is set to 1. The CRC code covers the start of frame, arbitration,

Fig. 15.11 CAN bus data frame

control, and data fields, using a CRC code suited for frames of less than 127 bits.

The ACK field comprises two bits: the ACK slot followed by the ACK delimiter. The transmitter sets both of these bits to 1. Any receiver which receives the message without error will set the ACK slot to 0, which indicates to the transmitter that the message has been received successfully.

The end of frame completes the frame with a sequence of seven bits set to 1.

An extended data frame is similar to the base data frame except that it contains an additional identifier field. The extended data frame has the identifier extension bit in the control field set to 1, an additional 18-bit identifier field, and an extra 'substitute' remote request bit which is set to 1. The extended data frame permits more address and message type information to be included within a message.

15.5.2.2 Remote Request Frames

A remote request frame is the same as a data frame except that the remote transmission request bit in the arbitration field is set to 1 to indicate that it is a remote request, and there is no data field.

To ensure that the receiver can maintain correct bit synchronization and to support error signaling, the bit sequences for data or remote frames are encoded using bit stuffing before they are transmitted. If there is a sequence of five consecutive bits of identical value, the transmitter will automatically insert an extra bit of the opposite value. This means that there will never normally be a sequence of six or more bits of the same value. Bit stuffing is applied to the start of frame, arbitration, control, data and CRC fields of the data or remote frames. The other fields are of fixed format and are not subject to bit stuffing.

15.5.2.3 Error Frames

If a node detects an error while receiving a data frame or remote frame, it will transmit an error frame. The error frame consists of an error flag and an error delimiter. If the node is in error active state, the error flag is an active error flag, which comprises six bits all set to 0, otherwise it is a passive error flag, which comprises six bits all set to 1. The error frame therefore violates the rule that no more than five

consecutive bits of the same value will be transmitted, or violates the fixed structure of the ACK field and end of frame. This means that all nodes on the bus will detect the fact that an error has occurred.

15.5.2.4 Overload Frames

The overload frame is similar to the error frame, signaling that at least one receiver needs some additional time before it will be ready to receive the next message.

15.5.2.5 Bus Arbitration Mechanism

If two or more nodes decide to start transmitting a data or remote frame at the same time there will be a clash of the two frames on the bus. To resolve this problem a CAN uses dominant (0) and recessive (1) bit values. A dominant value on the bus will override any recessive values on the bus, in a similar manner to a wired-AND, open drain operation. Each node is given a unique identifier, which is transmitted in the arbitration field. The first bit from both nodes will be the start of frame bit, which is 0 so that both nodes agree. As the next bits are sent, eventually one node will send a 1 and the other a 0. The node that sent the 0 wins (is dominant) and the node that tried to transmit a 1 but saw a 0 on the bus instead, will cease transmitting. Therefore the node with the lowest value identifier will win the arbitration and be permitted to send its message. Nodes that fail to win arbitration get another chance once the current frame has been transmitted. The CAN arbitration process is illustrated in Fig. 15.12.

The message identifiers must be unique in order to prevent two nodes with the same identifier sending messages that both get through arbitration but subsequently interfere with one another, resulting in an error. The allocation of identifiers can be done according to the sending node and the type of data being transferred. However, more efficient use of the bus bandwidth can be achieved by allocating the identifiers according to the deadline of the message. The shorter the deadline, the lower the identifier value, giving higher message priority.

Each node has its own clock for generating and receiving bits of CAN messages, which are nominally of the same frequency. The receivers all synchronize on the falling edge of the start of message field. An oversampling scheme is

Fig. 15.12 CAN bus arbitration process

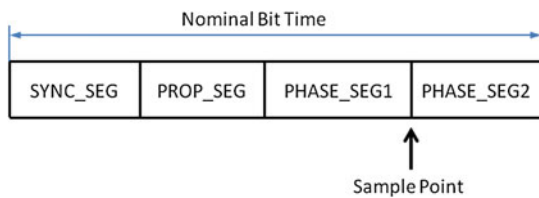
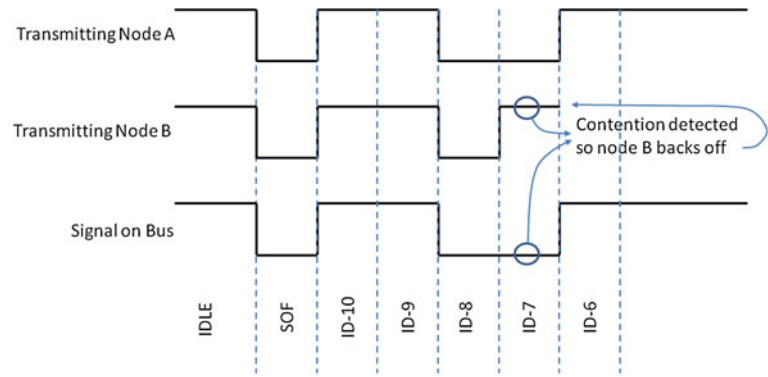


Fig. 15.13 CAN bus bit sampling

then used to determine where to sample subsequent bits as indicated in Fig. 15.13. Re-synchronization occurs every time there is a falling edge in the bit stream.

15.5.2.6 CANopen

CANopen [22] is a higher layer protocol for CAN which has been adopted and extended by ESA for spacecraft applications [23, 24].

The CANopen protocol stack is illustrated in Fig. 15.14. An object dictionary provides the interface between the application and the CAN network. Present in each CAN device, the object dictionary describes the functionality, parameters, and data pool of that device. The object dictionary is split into two parts: general specifications, which include a standardized device profile area and manufacturer specific information; and device specific specifications, which include the data area providing access to the data, and the data type information and the communications profile area, which define the communication mechanisms in use and related parameters.

The CANopen communications interface lies between the object dictionary and the CAN bus and provides a range of communications services.

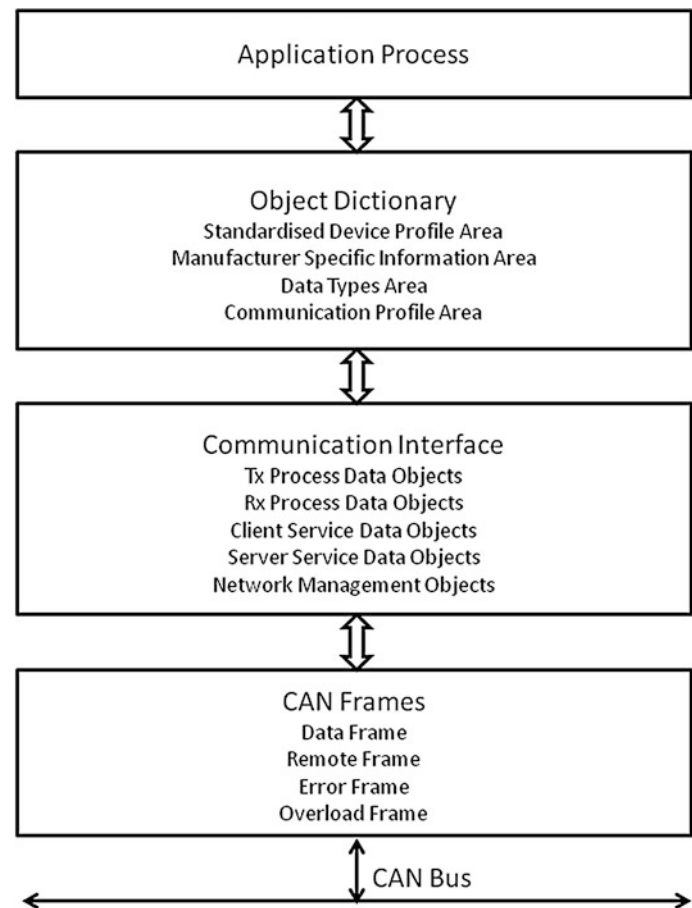
Process data objects (PDO) are used to communicate real-time information over the CAN bus. Transmit PDOs are for reading data from a device and receive PDOs are for sending data to a device. A PDO can transfer 8 bytes of data to or from a device, using the basic CAN bus communications facilities. A PDO can contain one or more object dictionary entries. The data objects to be transferred in a PDO are configurable via the mapping and parameter entries in the

communications profile area of the object dictionary. PDOs can be sent asynchronously, on request, or in response to an event occurring in the device; or synchronously following reception of a synchronization object (SYNC). Synchronous data transmission can be cyclic, after a specified number of SYNCs have been received; acyclic, on receipt of the SYNC following the occurrence of an event, and by request, on receipt of SYNC after a request to send data.

Service data objects (SDOs) are used for device configuration and sending larger amounts of data than can be sent in a single CAN frame. The SDO implements segmentation and de-segmentation of messages larger than the 8 byte limit of a CAN frame. An SDO uses a client-server communications model to read and write values in object dictionary in a remote node. The SDO client can have direct access to the object dictionary entries of another SDO server node. Two CAN identifiers are used to define one SDO channel, one for the client request and the other for the server response.

CANopen provides several network management functions

- Synchronization object (SYNC), used to synchronize the collection and transmission of data from various devices, and to provide high resolution time synchronization.
- Time stamp object (TIME), which contains time information in milliseconds and can be used together with SYNC to provide higher resolution time synchronization of nodes.
- Emergency object, which is a high priority message, used by a device to signal that an internal fatal error has occurred.
- Node control service, which is used to adjust the state of a node, allowing a network management node to cause a node to initialize, become operational or halt.
- Heartbeat service, where a node periodically signals its status.
- Node guarding service, where a network manager periodically polls every node to check that they are still in operation.
- Life guarding service, where a node will signal to its application if it has not been polled by the network manager within a specified period of time.

Fig. 15.14 CANopen protocols

15.5.2.7 CAN for Space

ESA have adopted CANopen as a higher layer protocol for the CAN bus, providing some extensions and recommendations to make it suitable for spacecraft applications [23, 24]. The extended identifier is used and a large data unit transfer (LDUT) protocol has been defined, which is recommended for segmented data transmission. A time distribution protocol has been added using dedicated PDOs to distribute spacecraft elapsed time (SCET) or universal time coordinated (UTC) time.

Recommendations are also made for bus architectures implementing redundancy.

Some manufacturers of spacecraft have implemented their own higher layer protocols [25].

15.5.3 SpaceWire

SpaceWire is a communications network designed specifically for use on-board spacecraft to connect together instruments, mass memory, processors, downlink telemetry, and other subsystems [26–28].

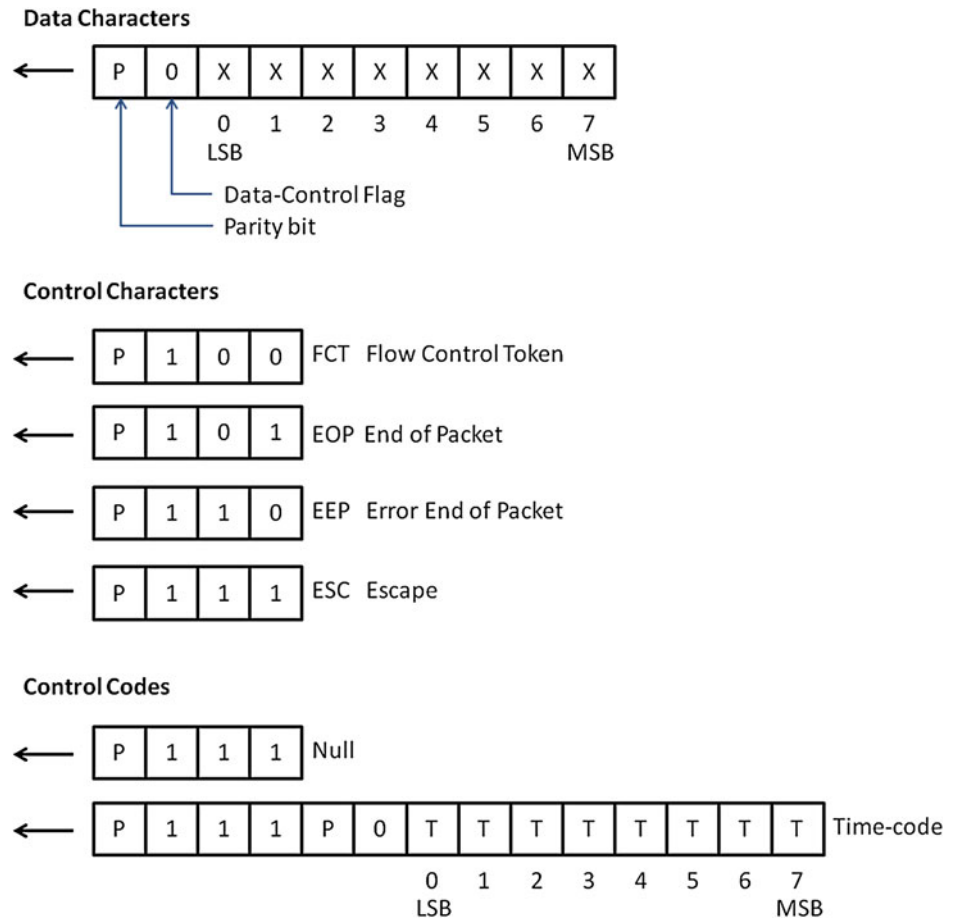
SpaceWire is simple to implement and has some specific characteristics that help it to support data-handling

applications in space: high-speed, low-power, simplicity, relatively low implementation cost, and a flexible architectural it adaptable to many space missions. SpaceWire provides high speed (2–200 Mbits/s), bidirectional, full duplex data links, which connect together SpaceWire enabled equipment. SpaceWire can be used to provide a direct connection between two units, for example, from an instrument to the mass memory unit. More complex data-handling networks can be built to suit particular applications using point-to-point data-links and routing switches.

SpaceWire is not galvanically isolated. Depending on the network topology and command mechanisms, SpaceWire may not be deterministic, so it cannot be used for command and control applications without due care. A deterministic data delivery protocol for SpaceWire is currently under development.

The SpaceWire standard was originally published in January 2003, and since then ESA, NASA and JAXA have used it for many scientific and Earth observation spacecraft. SpaceWire is also being used on several commercial spacecraft. High-profile missions using SpaceWire include: Gaia, Bepi-Colombo, Sentinel-1, 2-, -3 and -5-precursor, James Webb Space Telescope, GOES-R, Lunar Reconnaissance Orbiter, and Astro-H.

Fig. 15.15 SpaceWire data characters, control characters, and control codes



15.5.3.1 SpaceWireLinks

SpaceWire links are point-to-point data links that connect together a SpaceWire node (e.g. instrument, processor, mass memory unit) to another node or to a router. Information can be transferred over both directions of the link at the same time. Each link sends information as a serial bit stream using two signals in each direction (data and strobe). These signals are driven across the link using low voltage differential signaling (LVDS) [29] which requires two wires for each signal, resulting in a SpaceWire cable containing four screened twisted pairs.

15.5.3.2 SpaceWire Characters and Control Codes

The data characters, control characters, and control codes that can be sent over a SpaceWire link are illustrated in Fig. 15.15.

Each character starts with a parity bit that covers the data bits of the previous character and the parity and control bit of the current character. The parity bit is followed by a data/control flag, which is 0 for a data character and 1 for a control character. Data characters have eight data bits which contain a byte of data. Control characters have two bits which indicate the type of character, namely

- Flow control token, which is responsible for managing the flow of data across a link, to ensure that data is sent only when there is room for it at the other end of the link.
- End of packet (EOP) marker, which is used to signal the end of a SpaceWire packet.
- Error end of packet (EEP) marker, which is used to terminate a packet which has been cut off prematurely, because of an error on the link.
- Escape (ESC), which is used to construct control codes.

The two control codes are Null, which is ESC followed by an FCT, and the time-code, which is ESC followed by a data character.

The characters can be divided into three types

- Normal characters (N-Chars) used to form SpaceWire packets which are made up of data characters, EOPs, and EEPs. These characters are passed to/from the user application by a SpaceWire interface.
- Link control characters (L-Chars) used to initialise and maintain the running of a SpaceWire link, made up of Nulls and FCTs.
- Time-codes used to broadcast time or synchronisation information over a SpaceWire network.

Fig. 15.16 SpaceWire packet format

The SpaceWire characters are transmitted serially using data/strobe encoding to provide a clock signal for recovering the bits making up the characters at the other end of the SpaceWire link. Data/strobe encoding combines the serial bit clock and the data stream to form a strobe signal in such a way that XORing (XOR, 'Exclusive Or') the data and strobe signal recovers the clock signal. The data and strobe are then transmitted using LVDS. Data/strobe encoding is used instead of simply sending the bit clock along with the data, because it provides better skew tolerance.

Each SpaceWire interface contains a link state-machine which is responsible for starting a link, keeping the link running, sending data over the link, ensuring that data is not sent if the receiver is not ready for it, and recovering from any errors on the link. The operation of the link state-machine is transparent to the user application that wishes to send data over the link.

15.5.3.3 SpaceWire Packets

Information is transferred across a SpaceWire link in distinct packets. Packets can be sent in either direction of the link, provided that there is room in the receiver for more data. The SpaceWire packet is formatted as illustrated in Fig. 15.16.

The destination address is a list of data characters that represents either the identity of the destination node or the path that the packet has to take through a SpaceWire network to reach to the destination node. In the case of a point-to-point link directly between two nodes (no routers in between) the destination address is not necessary.

The cargo is the data to be transferred from source to destination. Any number of data bytes can be transferred in the cargo of a SpaceWire packet.

The data character following the EOP marker is the start of the next packet. There is no limit on the size of a SpaceWire packet.

SpaceWire packets can be used to carry a range of user protocols, with minimal overhead.

15.5.3.4 SpaceWire Networks

SpaceWire networks are constructed using SpaceWire point-to-point links and routing switches. A SpaceWire router [27] connects together many nodes using SpaceWire links, providing a means of routing packets from one node to any of the other nodes or routers attached to the router. A node is simply the source or destination of a SpaceWire packet. A SpaceWire router comprises a number of SpaceWire link interfaces and a switch matrix. The switch

matrix enables packets arriving at one link interface to be transferred to and sent out on another link interface on the router.

Each link interface may be considered as comprising an input port (the link interface receiver) and an output port (the link interface transmitter). A SpaceWire router transfers packets from the input port of the routing switch where the packet arrives, to a particular output port determined by the packet destination address. A router uses the leading data character of a packet (one of the destination address characters) to determine the output port of the router to which the packet is to be forwarded. The destination address is the first part of the packet to be sent so that a router, which is responsible for forwarding the packet along the next link towards its destination, only has to receive the first character of a packet before making the decision about where the packet is to be routed. This makes routing very fast provided no other packet is currently using the link that the packet is to be forwarded along. If there are two input ports waiting to use a particular output port, when the previous packet has finished being sent an arbitration mechanism inside the router decides which of the waiting input ports is to be served next. The other one has to wait until the output port becomes free once more.

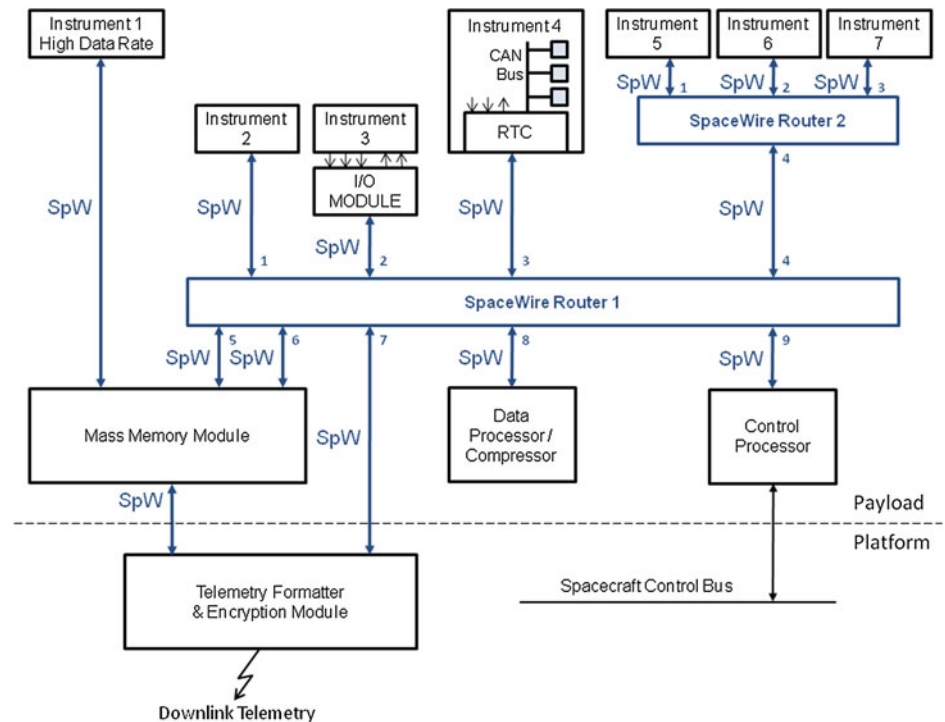
SpaceWire is able to support many different payload processing architectures using point-to-point links and routing switches. The data-handling architecture can be constructed to suit the requirements of a specific mission, rather than having to force the application onto a restricted bus or network with restricted topology.

15.5.3.5 Packet Addressing

The destination address at the front of a SpaceWire packet is used to route the packet through a network from the source node to the destination. There are two forms of addressing used in SpaceWire networks: path addressing and logical addressing.

Path addressing has one byte in the destination address for each router on the path from the source to the destination. When the packet arrives at a router, the router takes off the first byte of the packet and uses it to determine which port to forward the rest of the packet through. For example, if the leading byte is 3 the packet is forwarded through port 3 of the router. If the packet header is 3, 2, 4 the packet will be routed through port 3 at the first router it encounters, through port 2 at the next one, and finally through port 4 at the last router before it reaches its destination. At each router the leading character is used to determine where to

Fig. 15.17 Example SpaceWire architecture



forward the packet and is then discarded to reveal the next address character for use at the next router. A router has an internal configuration port (port 0) and up to 31 external ports (1–31), so a path address byte is always in the range 0–31.

Logical addressing uses the remaining values (32–255) of the leading byte of a packet as an index into a routing table. Each node on the network is given a unique logical address in this range. To send a packet to a node using logical addressing, the destination address of the packet contains one byte which is set to the logical address of the destination node. When the packet arrives at a router the router recognises that the leading byte is a logical address because it is in the range 32–255. It reads the value of the logical address and uses it to look up the output port number that it should forward the packet through to send it towards its destination. Each router has a routing table programmed with an entry for each logical address to indicate the output port that the packet should be forwarded through. In the case of logical addressing the first address character is not deleted after it has been used by the router, but is left at the front of the packet so that it can be used again at the next router.

A path address determines the path that a packet will take through the network, with an explicit instruction about which way to go for each router on that path. A logical address specifies the destination that the packet is intended for, and it is up to the routers with their routing tables to sort out which output port to send the packet through.

15.5.3.6 Example SpaceWire Architecture

An example SpaceWire architecture is shown in Fig. 15.17, for a typical payload data system. It uses two SpaceWire routers to provide the interconnectivity between instruments, memory, and processing modules.

Instrument 1 uses a SpaceWire point-to-point link to stream data directly into the mass memory module. If the data rate of the instrument is greater than that of a single SpaceWire link, two or more links may be used in parallel.

Instrument 2 is connected to the mass memory module via router 1. This configuration allows it to send data to other units like the data processor/compressor, and to receive commands from the control processor.

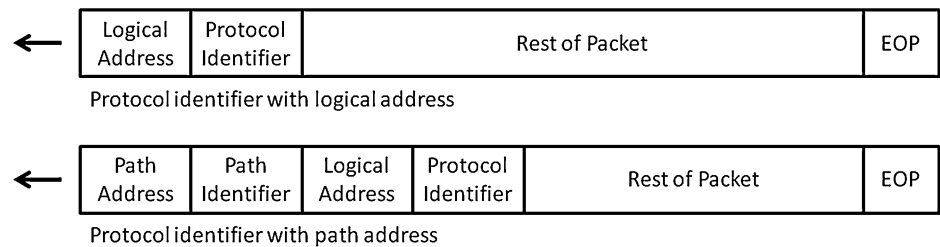
Instrument 3 does not have a SpaceWire interface so an input/output (I/O) module is used to connect the instrument into the SpaceWire network.

Instrument 4 is a complex instrument containing a number of sub-modules which are interconnected using the CAN bus. A remote terminal controller (RTC) is used to bridge between the CAN bus and SpaceWire.

Instruments 5, 6 and 7 are located in a remote part of the spacecraft. To avoid having three SpaceWire cables running to this remote location a second router (SpaceWire Router 2) is used to concentrate the information from these three instruments and to then connect to the main router.

This mass memory module can receive data from any of the instruments either directly, as is the case for Instrument 1, or indirectly via Router 1. Data stored in the mass memory

Fig. 15.18 SpaceWire protocol identifier



module can be sent to the telemetry formatter/encryption module for sending to Earth, or it may first be sent to a data processing or data compression unit. This unit may return the processed/compressed data to the mass memory module or send it straight to the telemetry module via Router 1.

The control processor is able to control all the Instruments, mass memory module and telemetry units via the SpaceWire network. It can configure, control and read housekeeping and status information from them.

In Fig. 15.17 redundancy has been eschewed for clarity. In a space flight application, an additional pair of routers would be included with duplicate links to the modules to provide redundancy. It is straightforward to support traditional cross-strapped, redundant modules using SpaceWire.

15.5.3.7 SpaceWire Time Codes

SpaceWire time codes [28] provide a means of sending time or synchronization information across a SpaceWire system. Time information can be provided as ‘ticks’ or as an incrementing value which may be synchronised to spacecraft time. The time codes are broadcast rapidly over the SpaceWire network, alleviating the possible need for a separate time distribution network. A time code contains an 8-bit value, 2 bits of which are flags set to 0b00 and the remaining 6 bits are incremented each time a new time code is broadcast. When a time code is received in a node or a router its value is checked against a local time code counter. If the incoming time-code is one more than the value of the time code counter, it is accepted and passed on to the application. In the case of a router, accepted time codes are forwarded out through all active links of the router except the one from which the time code originated. If the incoming time code is not one more than the value in the local time-code counter, it is rejected. Whether the incoming time code is valid or not, its value will be loaded into the local time-code counter after comparison with the counter value. This mechanism provides a simple means of broadcasting the time code over a network with arbitrary structure.

15.5.3.8 SpaceWire Remote Memory Access Protocol

A simple, but powerful packet structure allows SpaceWire to carry other protocols and be used for many different applications. A set of protocols that run over SpaceWire

have been devised to support some specific needs spacecraft on-board communications.

To carry a set of protocols over SpaceWire it is necessary to be able to distinguish one protocol from another. This is done using a specific packet format which contains a protocol identifier, as illustrated in Fig. 15.18.

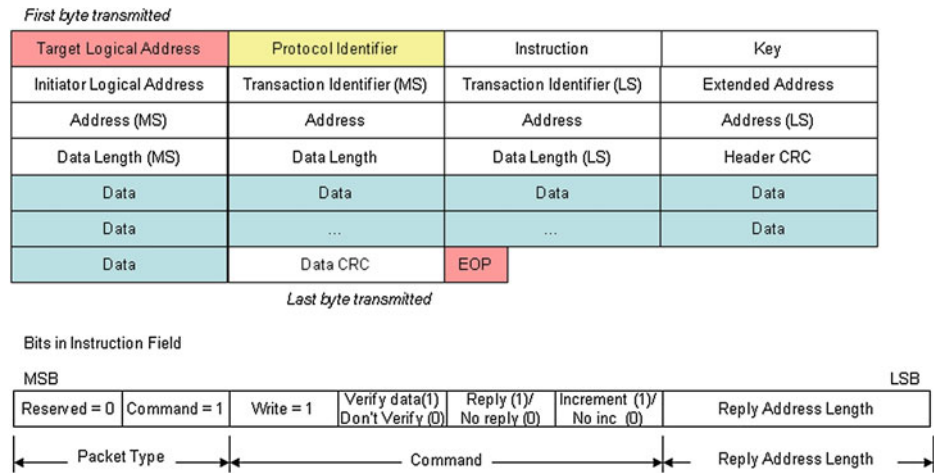
The protocol identifier is a single data character which identifies the specific protocol being carried by the SpaceWire packet. When logical addressing is being used the protocol identifier follows the logical address. When path addressing is being used a dummy logical address is added following the path address characters, and the protocol identifier follows that dummy logical address. The dummy logical address can be an actual logical address of a destination node or the default logical address of 254. The path address bytes are stripped off as the packet travels through the network, exposing the dummy logical address and protocol identifier in the same positions as when logical addressing is being used. This allows the position of the protocol identifier to be determined regardless of which addressing mode is being used.

One of the protocols that have been standardized by ECSS is the SpaceWire remote memory access protocol (RMAP) [30].

The RMAP provides a common mechanism for reading and writing to registers and memory in a remote device over a SpaceWire network. It can be used to configure devices, read housekeeping information, read data from an instrument or mass memory, and write data into a mass memory from an instrument. Together RMAP and SpaceWire provide a powerful combination for spacecraft instrument data-handling. A central payload processing computer is able to configure the instruments using RMAP. When data is available it can be read from the instrument using an RMAP read command. A uniform memory space can be provided for each instrument with pages for instrument data, configuration registers and housekeeping status registers, simplifying and standardizing instrument control operations.

Figure 15.19 shows an RMAP write command being sent when logical addressing is being used. The node that is sending the command is referred to as the initiator, and the node receiving the command is the target. The RMAP command begins with the logical address of the target node followed by the protocol identifier, which is 1 for RMAP.

Fig. 15.19 SpaceWire RMAP write command format



The next data character is the RMAP instruction, which is detailed in Fig. 15.19. This data character indicates whether the RMAP packet is a command or a reply, the type of the command (write, read, or read-modify-write), whether for a write command the data is to be verified before it is written, if a reply is required, and if the address being written is to be incremented as data is written to it. The reply address length is used when RMAP is being sent with path addresses.

The key is a data character whose value has to be agreed between initiator and target for the command to be accepted. The initiator logical address is the logical address of the initiator and used when sending a reply.

The transaction identifier is a 16-bit field that is normally incremented each time an RMAP command is sent out by the initiator. Its value is included in the reply and allows the initiator to associate the reply that it receives with the command that caused the reply.

The extended address and address fields for a 40-bit address which specifies where the data in the write command is to be written. This may be either a physical or virtual address.

The data length specifies the amount of data included in the write command.

The header CRC is used to check that the header does not contain any errors before it is acted upon. The target node will also check that the various fields in the header are acceptable before it authorizes the write operation. For example, it may reject commands from initiators that ought not to be writing data to it, or it may reject attempts to write to certain areas of memory.

The data field contains the data to be written to memory in the target node.

The data CRC is used to validate that the data written did not contain any errors. For small amounts of data it is possible to buffer it and check it using the data CRC, before it is written into memory. This is requested by setting the

verify data bit in the instruction field of the RMAP command. Otherwise, data will be written directly into memory and the data CRC used to indicate whether there were any errors.

RMAP supports write, read, and read-modify-write commands, and also allows commands to be posted, i.e. many commands to be sent without waiting for a reply to be received before sending the next command.

RMAP is currently being used in several European, American, and Japanese spacecraft.

15.6 Recent On-Board Data Storage Technologies

The technology used for storage of large amounts of data on-board spacecraft has varied over the years. Tape recorders were the natural choice in the early days despite their relatively low reliability, where many failures were due to malfunctions of the mechanical tape transport. Tape recorder data management was also problematic, as a playback operation had to be preceded by a rewind or the data had to be replayed backwards. The tape recorder playback mechanisms even affected early telemetry protocol standardization [31]. Since the 1990s the tape recorders have been superseded by solid-state recorders (SSR) in almost all applications. Over the years, there have also been attempts to employ other mass storage technologies like magnetic bubble memories and hard disks, but with limited success.

The early SSRs were based on dynamic random access memory (DRAM) technology and had capacities of the order of one Gbit. DRAM technology is by its nature volatile. A DRAM memory needs continuous refreshing in order not to lose data. This is done by regularly recharging its internal memory cells consisting of small capacitors. Consequently, if power is lost the memory content is lost.

SSRs are currently dominated by three memory technologies, SDRAM, DDRAM, and Flash.

15.6.1 Single Data Rate Memories

The first technology is single data rate (SDR) synchronous DRAM (SDRAM). The technology is known from Pentium4-based PCs where the memory modules were called 'PC-100' or 'PC-133', the number being the clock rate used for communication with the memories. There were several versions of the SDR SDRAMs that were suitable for space applications, and the space industry has acquired a significant stock that is still used when producing new units of existing designs. The production of these memories has now ceased, and for new designs other memory types are used. The SDR technology produced memory chips that had a maximum size of 512 Mibits (where 1 Mibit = 2^{20} bits), typically arranged as $128\text{ M} \times 4$ bits or $64\text{ M} \times 8$ bits.

The main problems to solve when using SDR memories in space applications are

- Tolerance to multiple bit upsets due to single event effects (SEE). An ion that hits a memory chip may change multiple bits of the chip. The multiple bits may affect a single word or multiple words depending on the chip architecture. The solution to the problem is to use an error correcting code that can correct any error that can be present in single chip. The most common method is to use a Reed-Solomon code that has a symbol width equal to the size of the memory chips used. If 8-bit wide chips are used in a 32-bit wide memory, the typical code to use would be an RS(6,4) code that can correct any byte error in a 48-bit memory word. The error coding overhead would for this case be 50 % since 16 bits of additional information is needed for each 32-bit word in the mass memory.
- Tolerance to single event functional interrupts (SEFI). An ion that hits a memory chip may cause the entire chip to malfunction. There are only two ways to solve the problem. The first method is to power down and re-power the failing chip. Another method is to reinitiate the failing chip using the chip specific control commands. In both cases the same error correcting code, e.g. an RS(6,4) code, should be sufficient to recover the proper data stored.

The main advantages of the SDR memories are the relatively simple interface in combination with the 3.3 V power supply. This allows implementing an SDR memory in most ASIC and field programmable gate array (FPGA) technologies and it also allows designing parallel data buses that can be configured to interface both SDR memories and SRAM memories.

15.6.2 Double Data Rate Memories

The second technology used is double data rate (DDR) SDRAMs. The name originates from the fact that DDR memories are capable of transferring data on both the rising and the falling edges of the clock, contrary to the SDR memories that transfer data on a single clock edge only. DDR memories have seen five commercial generations, where DDR-3 is now the dominating memory for PC motherboards and DDR-5 is used extensively in modern high-end graphics boards. Just as for the SDR memories, there are some DDR memories that can be used in space applications. The main drawback with various DDR generations is that they are not compatible with each other. Even if both DDR-2 and DDR-3 memories use both edges of the clock, they differ in for instance signaling voltage levels, timing, and pre-fetch buffer size. The DDR-3 chips are available with a maximum size of 1,024 Mibit, while DDR-5 chips are available with a maximum size of 2,048 Mibit.

All SDRAM memories are accessed in burst mode, where the maximum burst length varies between the generations. Burst accessing means high average data bandwidth but with an increased delay, called latency, from initiating the access to the data becoming available. To operate efficiently with general purpose processors, it is necessary that the processor have a cache memory whose line size matches the memory burst size. Efficient operation in mass memory applications is easier, since the data access is, by its nature, more sequential compared to operating with a processor. However, there must be some data buffering in the memory module interface in order to handle the burst accesses, and to handle the memory refresh and memory scrubbing operations that take place at regular intervals.

The DDR memories have the same error behavior as the SDR memories, i.e. multiple bit errors or SEFI errors may occur as the result of a single ion hitting the chip.

15.6.3 Flash Memories

The main disadvantage of today's SDRAM devices is the limited storage capability compared to the third technology used in spacecraft mass memories, the Flash memory. The Flash memory technology is non-volatile, meaning that the memories keep their content even if powered off. There are NOR and NAND types, and they differ in the way they are accessed. NOR Flash memories can be accessed by random addresses and are typically used to store program code for processors. NAND Flash memories are accessed sequentially and are typically used as mass data storage devices in USB sticks, cameras, and solid-state drives. Commercial

NAND Flash memory technology is today at 64+ Gbit per memory chip. This is a factor of 32 better than the DDR-5 technology, and consequently NAND Flash memory technology is better when implementing large mass memories. NOR Flash memories have lower densities and are today at sizes similar to SDR memories. Thus, the NOR flash memories have gained less interest in space applications.

NAND Flash memories are arranged in pages and blocks. A typical page size is 4,096 + 128 bytes and a typical block consists of 128 pages. Accesses must be done page by page and a page can only be written if the block that it belongs to has been erased since the page was last written.

The main limitation with the current NAND Flash memories is the access bandwidth. A typical memory chip of 8 bits can be written with a rate of up to 40 Mbytes per second but when one page is written a programming cycle must take place. The programming cycle typically lasts ten times longer than the page writing time and during the programming the chip is basically unavailable for other accesses. Thus, the programming cycle effectively limits the input data rate to about 4 Mbytes per second. The main way to handle the writing limitations is to write several memory chips in parallel. During the programming of one chip it is possible to write to several other chips in parallel, and the programming cycles of the different chips can be staggered such that one chip is always available for writing.

The second constraint is a limited number of erase and programming cycles. A block can only be erased about 100,000 times and may require that some mechanism be implemented to ensure that the limit is not exceeded over the mission lifetime.

The third limitation is that the memory chips may contain faults when delivered from the supplier. These faults are marked at chip delivery, and the user must create and maintain a bad block list per chip somewhere in his system. The main failure mode during operation is that individual bits may fail and the bad block list must then be updated to prevent that faulty block from being used.

Failing bits can be corrected at read-out by applying various error correction schemes. All error correction schemes need extra information bits, and this has already been foreseen in the NAND Flash chips where each page has sufficient extra bytes to store the error correcting code. Flash chip manufacturer publish suitable algorithms to use for their products.

15.7 Future On-Board Data Systems Technologies

This section presents the state at the time of writing. Since research and technology is moving fast, some information may be obsolete or changed at the time of reading.

15.7.1 Multi-Core CPUs

Commercial processors today are more and more based on the use of several CPU cores embedded in a single chip. Dual core processors for space applications are announced by Aeroflex Gaisler (GR712RC) and Space Micro (Proton400k-L). The Aeroflex Gaisler solution uses a dual LEON3 and is implemented on a 0.18 μ process with the CPU running at up to 125 MHz. The Space Micro solution uses a dual PowerPC e500 and is implemented in a 45 nm Silicon-On-Insulator (SOI) process with the CPU running between 800 MHz and 1.2 GHz. The Space Micro solution exploits commercial technology like the Maxwell solution described in Sect. 15.4.2 but uses a different concept to achieve the desired SEU tolerance. While the Maxwell solution uses three parallel operating processors and majority voting, the Space Micro solution uses a single processor and compares multiple serial execution sequences. A special technique has been developed that exploits the fact that modern processors may have individually controllable parallel execution units whose results can be compared. At least two execution sequences are necessary to determine whether an error has occurred, and by executing a third sequence it is possible to determine the correct result.

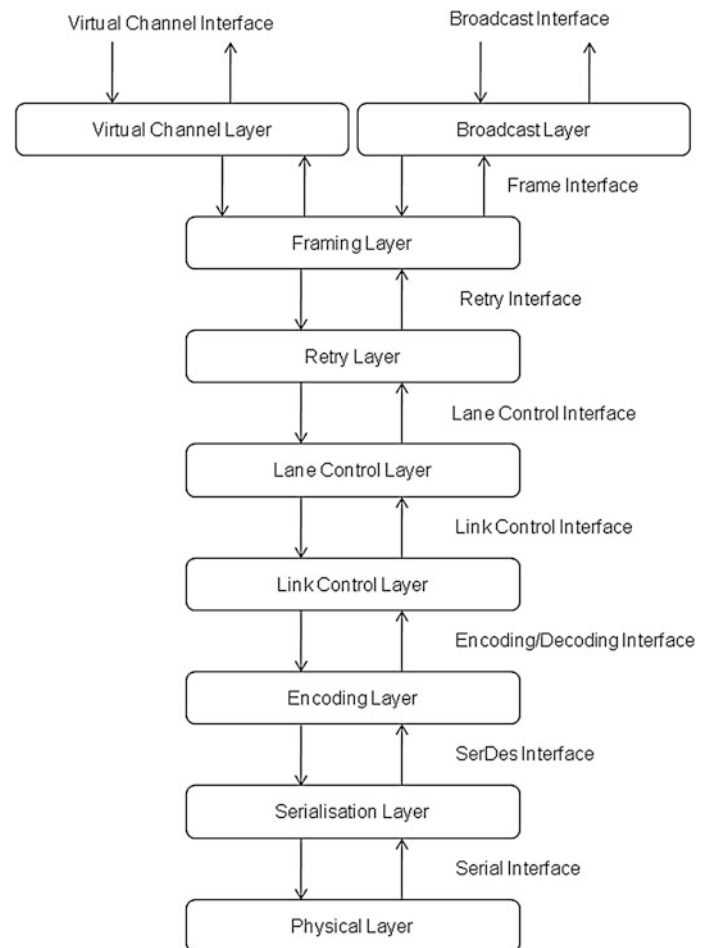
A quad core microprocessor called Next Generation Microprocessor (NGMP) is being developed under an ESA contract [32]. It is based on four LEON4-FT CPUs and includes a PCI port and a number of fast serial links. The target manufacturing process is a deep submicron technology that, depending on the final technology chosen, would allow CPU operation at up to 400 MHz.

15.7.2 SpaceFibre

Several instruments, including synthetic aperture radar and multispectral imagers, require higher data rates to the mass memory unit than can be provided by SpaceWire. Downlink telemetry systems are being designed that can support Gbit/s data transfer rates leading to the need for similar data rates to transfer the data from the mass memory unit. There is a growing requirement for a data communications link with an order of magnitude higher performance than SpaceWire, which can be implemented in radiation tolerant, space-qualified technologies.

SpaceFibre is a very high-speed serial communications link which is being designed for use on spacecraft [33]. A SpaceFibre link connects high data rate payloads into the on-board data handling system and is able to interoperate with a SpaceWire network. A SpaceFibre link can operate over a copper or fibre-optic communications medium and support real data rates of more than 2 Gbit/s, an improvement over SpaceWire of at least a factor of 10.

Fig. 15.20 SpaceFibre CODEC architecture overview



SpaceFibre sends and receives SpaceWire packets and is able to send many SpaceWire packets at the same time over a single SpaceFibre link using virtual channels. Each SpaceWire packet is split up into small frames to send over the link, so that information from many virtual channels can be multiplexed over the link. A medium access controller determines which output virtual channel buffer (VCB) is allowed to send data over the SpaceFibre link. SpaceFibre provides several quality of service levels

- Priority, where the virtual channel with the highest priority sends the next frame.
- Bandwidth reserved, where the virtual channel with allocated bandwidth and recent low utilisation of the link sends the next frame.
- Scheduled, where virtual channels are allocated specific time-slots in which to send frames.

SpaceFibre virtual channels provide the capability for high data-rate communications on-board a spacecraft, with a quality of service suitable for many spacecraft data-handling applications. SpaceFibre also has a low latency message broadcasting mechanism which can be used to distribute time, signal events, and send other short messages over the SpaceFibre network (Fig. 5.20).

15.7.3 Other Future Developments

Several terrestrial technologies are likely to have a significant impact on spacecraft on-board processing and data-handling in future.

System-on-chip technology is already being used in spacecraft applications resulting in substantial reductions in the mass, size, and power consumption of data system computers. The mass and power of the pure computer part was reduced, by the order of, a factor of two in the first decade of the 21st century. This trend in system-on-chip integration is likely to continue and result in standard processing chips that integrate much of the on-board processing functionality in a single device, including processors, memory, interfaces, and combined with FPGAs for hardware customisation. However, evolution in other technology areas such as power converters, communication links and packaging technology is required to achieve significant further mass and power reductions.

Reconfigurable FPGA technology offers the potential for high-speed data processing, which can be reconfigured to perform different processing functions at different stages of a mission or for different instruments. Wireless communication

technology offers significant potential for mass reduction by removing the need for data network cabling.

Technology spin-in from terrestrial applications is likely to spur on the development of spacecraft on-board data-handling systems to enable higher data-rate instruments to be supported with more demanding processing, compression, and data storage requirements. At the same time, technologies developed for spacecraft applications with their particular requirements for reliability and simplicity are being spun-out into niche terrestrial applications.

References

1. CCSDS 130.0-G-2: Overview of Space Communications Protocols. Green Book. Issue 2. December 2007. Available at www.ccsds.org.
2. CCSDS 200.0-G-6: Telecommand Summary of Concept and Rationale. Green Book. Issue 6. January 1987.
3. CCSDS 231.0-B-2: Telecommand Synchronization and Channel Coding. Blue Book. Issue 2. September 2010
4. CCSDS 232.0-B-2: Telecommand Space Data Link Protocol. Blue Book. Issue 2. September 2010
5. CCSDS 232.1-B-2: Communications Operation Procedure-1. Blue Book. Issue 2. September 2010
6. CCSDS 133.1-B-2: Encapsulation Service. Blue Book. Issue 2. October 2009
7. CCSDS 133.0-B-1: Space Packet Protocol. Blue Book. Issue 1. September 2003
8. CCSDS 100.0-G-1: Telemetry Summary of Concept and Rationale. Green Book. Issue 1. December 1987
9. CCSDS 131.0-B-1: Telemetry Synchronization and Channel Coding. Blue Book. Issue 1. September 2003
10. CCSDS 132.0-B-1: Telemetry Space Data Link Protocol. Blue Book. Issue 1. September 2003
11. ECSS-E-ST-50-14C: Spacecraft discrete interfaces, 31 July 2008
12. ECSS-E-70-41A: Ground systems and operations - Telemetry and telecommand packet utilization, 30 January 2003
13. CCSDS 301.0-B-4: Time Code Formats. Blue Book. Issue 4. November 2010
14. CCSDS 727.0-B-4: CCSDS File Delivery Protocol (CFDP). Blue Book. Issue 4. January 2007
15. J.C. Laprie, Dependability: Basic Concepts and Terminology Springer-Verlag, 1992. ISBN 0387822968
16. The SPARC Architecture Manual, Version 8, www.sparc.org
17. MIL-STD-1553B: Digital Time Division Command/Response multiplex data bus, Notice 2, 8 September 1986
18. MIL-HDBK-1553A: Multiplex applications handbook, Issue A, 1 November 1988
19. ECSS-E-ST-50-13C: Interface and communication protocol for MIL-STD-1553B data bus onboard spacecraft. Issue C, 15 November 2008
20. Bosch "Controller Area Network (CAN) Specification", Version 2.0, Robert Bosch GmbH, 1991.
21. "Road vehicles – Interchange of digital information – Controller Area Network (CAN) for high-speed communication", International Organisation for Standardisation, Geneva Switzerland, ISO11898:1995 Amendment 1 (Amended 1995).
22. CANopen Application Layer and Communication Profile. CiA Standard 301, Version 4.01, CAN in Automation. Available from www.can-cia.de, last viewed 10/10/2011.
23. ECSS-E-ST-50-15C draft
24. C. Plummer, P. Roos, & L. Stagnaro, "CAN Bus as a Spacecraft Onboard Bus", Proceedings of DASIA 2003 (ESA SP-532). 2-6 June 2003, Prague, Czech Republic, p.51.1.
25. A. Woodroffe and P. Madle, "Application and experience of CAN as a low cost OBDH bus system", MAPLD 2004, Paper number p106, available from http://klabs.org/mapld04/program_sessions/session_p.html
26. ECSS Standard ECSS-E-ST-50-12C, "SpaceWire, Links, Nodes, Routers and Networks", European Cooperation for Space Data Standardization, July 2008.
27. S.M. Parkes, C. McClements, G. Kempf, S. Fischer and A. Leon, "SpaceWire Router", in: ISWS International SpaceWire Seminar 2003 (Noordwijk, The Netherlands, 4-5 November (CD Rom) 2003) pp.180-187.
28. S.M. Parkes, "The operation and uses of the SpaceWire time-code", in: ISWS International SpaceWire Seminar 2003 (Noordwijk, The Netherlands, 4-5 November (CD Rom) 2003) pp.223-230.
29. S.M. Parkes, "High-Speed, Low-Power, Excellent EMC: LVDS for On-Board Data-handling", DSP'98, 6th International Workshop on Digital Signal Processing Techniques for Space Applications, ESTEC, Noordwijk, The Netherlands, 23-25 September 1998, European Space Agency (ESA) publication no. WPP-144, paper P16
30. ECSS-E-ST-50-52C, SpaceWire—Remote memory access protocol, 5 February 2010
31. Andersson, J., Gaisler, J., Weigand, R., "Next Generation Multipurpose Microprocessor", DASIA, Budapest, Hungary, 2010
32. S.M. Parkes, C. McClements, M. Dunstan and M. Suess, "SpaceFibre: Gbit/s Links For Use On board Spacecraft", International Astronautical Congress, Daejeon, Korea, 2009, paper IAC-09-B2.5.8.
33. S.M. Parkes, "SpaceFibre Standard Draft D", February 2012, available from <http://spacewire.esa.int/WG/SpaceWire>

Christopher Krupiarz, Annette Mirantes, Doug Reid, Adrian Hill
and Roger Ward

As with other subsystems, improved hardware has allowed spacecraft software to become a rapid growth area within mission architectures. Early spacecraft either relied heavily on user commands from the ground or preprogrammed sequences to complete their objectives. As missions became more complex and computing hardware became smaller and more manageable, mission designers moved functionality from the ground systems to the flight computers. Early spacecraft computers were particularly resource constrained, requiring software developers to be heavily focused on full use of the limited capabilities of early central processing units (CPU). As flight computer hardware has advanced, so has software, to the extent that it is now an integral part of spacecraft systems.

16.1 Spacecraft Flight Software History/Evolution

16.1.1 Computing for Human-Rated Spacecraft

A review of National Aeronautics and Space Administration (NASA) human space flight computing provides an example of the general evolution of software in this area. On-board computing was first used in the 1960s. On the Gemini missions, software used assembly language, which closely matches the machine code used to execute computer instructions, to develop software to perform various spacecraft activities. For the Apollo program, NASA funded development of the

Apollo guidance computer (AGC) [1]. As with the Gemini computer, software developers used assembly language to create programs for the AGC. However, the AGC provided an additional programming capability called the Interpreter. Instead of relying solely on assembly, developers wrote software in a higher-level language that enabled more complex calculations. Although the programs ran slower than those written strictly in assembly, the interpreted code was less prone to errors and allowed for more extensive review and analysis. The AGC also used a real-time operating system (RTOS) to control the priority and scheduling of execution of its programs. Although greatly simplified compared to modern RTOSs, the AGC's operating system, called the Executive, enabled up to eight applications to run on the spacecraft at any given time. With the completion of the Apollo program, NASA again tasked IBM to develop the computing architecture for the Space Shuttle. Unlike Apollo and Gemini, the Shuttle computers were general purpose and used a higher-level language called High-order Assembly Language/Shuttle (HAL/S).

16.1.2 Computing for Robotic Spacecraft

As with human-rated spacecraft, on-board computing for robotic spacecraft began in the 1960s. NASA's early Moon missions, Ranger and Surveyor, used sequencers with a pre-defined set of instructions to account for all operations on a given mission, along with the ability to activate sequences upon command [2]. The Mariner 6 and 7 probes to Mars, developed by the NASA Jet Propulsion Laboratory and launched in 1969, saw the first capability to program a spacecraft during a mission. In Europe, the first mission to be controlled by software was the European X-ray Observatory Satellite (Exosat) Earth observation satellite, which was launched in 1983. As with early human spacecraft, engineers wrote the flight software for these spacecraft in assembly language.

C. Krupiarz (✉) · A. Mirantes · D. Reid · A. Hill
Embedded Applications Group Space Department, The Johns
Hopkins University Applied Physics Laboratory, Laurel, MD,
USA
e-mail: Christopher.Krupiarz@jhuapl.edu

R. Ward
SciSys Ltd., Bristol, England

As the complexity of deep-space missions increased, so did the computing power and capability. Deep-space probes led the way in the development of autonomous software because mission constraints required the spacecraft to make critical decisions at distances that made real-time human intervention impossible. Autonomous fault protection software was particularly important to ensure the safety of the spacecraft by responding to failures. This area grew in capability and complexity from the two Voyager spacecraft on through Galileo, Magellan, and Cassini. Galileo was the first deep-space probe to use the HAL/S programming language. As on-board spacecraft flight computers and microprocessors became faster and provided much greater memory storage, spacecraft flight software grew to fill the memory and to provide greater capabilities. Additionally, software engineers began to develop programs in standard higher-level languages such as Ada, C, and C++, thus matching the need for more complex software.

Although software engineers currently typically hand-code software, the emerging generation of flight software sees more auto-generated code, thus enabling developers to focus on a higher-level abstraction. This is already seen in areas such as guidance, navigation, and control (GN&C) where flight algorithms are generated with off-the-shelf tools. The next step is to apply these techniques to non-GN&C software as well.

16.2 Where is Flight Software Used?

In a modern spacecraft, software is an integral element of many spacecraft subsystems. On the main computer, software is generally categorized into three groups: (1) boot, (2) command and data handling (C&DH) and (3) GN&C.¹ In addition, software can be found in subsystems such as power and communication.

16.2.1 Boot

Spacecraft flight software typically runs on a single board computer (SBC) with a processor, on-board hardware, and peripherals. The boot software configures the SBC to run the application software. This includes initializing registers on the processor, preparing memory, and configuration of the peripheral devices.

¹ The definitions of these terms can vary from organization to organization or between application environments. For instance, for United States Department of Defense missions, telemetry, tracking, and command (TT&C) is used in place of C&DH.

16.2.2 Command and Data Handling

The C&DH subsystem controls the flow of data through the spacecraft. C&DH provides input to the spacecraft via command processing and output in the form of telemetry frames for mission operations personnel. The C&DH software also handles routing of data to various other subsystems either internal to the spacecraft computer or externally via a hardware communications bus. Other software functionality in the C&DH subsystem can include recording data to and playing data back from a solid-state recorder (SSR), performing fault management operations to ensure the health and safety of the spacecraft, and compressing data to reduce downlink bandwidth.

16.2.3 Guidance, Navigation, and Control

The primary functions of the GN&C subsystems are to maintain spacecraft attitude, to execute propulsive maneuvers for spacecraft trajectory control, and to provide a navigation function that maintains positional knowledge within a given frame of reference. Because GN&C is a control system, the execution of the GN&C software algorithms is tightly controlled. As such, they are usually executed at a fixed and known rate. Typical execution rates include 20 Hz for control and 1 Hz for guidance and navigation, which is correspondingly synchronized to mission elapsed time (MET), coordinated universal time (UTC), or some other time standard. The software ingests data from the GN&C sensors through coordination with the C&DH subsystem, runs the data through either hand-coded or auto-generated models, and outputs actions destined for actuators. As discussed in Chap. 12, typical GN&C hardware sensors and the corresponding rates include 100-Hz inertial measurement units (IMU), 10-Hz star trackers (ST), and 1-Hz Sun sensors (SS). Typical actuators include 20-Hz reaction wheels, 5-Hz magnetotorquer rods, and 20-Hz attitude control thrusters.

16.2.4 Payload Software

A standard spacecraft configuration consists of a spacecraft bus plus one or more payloads or instruments. Generally, instruments have a dedicated processor, with the flight software on this processor integrated into the overall architecture of the instrument, payload, or subsystem; the processor is responsible for operating the instrument as well as receiving and transmitting data from and to the main spacecraft computer. As instruments become more sophisticated, the complexity of this software is also rapidly increasing and can include significant data processing to reduce the amount of data to be downloaded.

16.2.5 Other Subsystems

Although the C&DH, GN&C, and payload constitute most of the software on a spacecraft, software is found in other subsystems as well. For example, a power system may contain software to control the rate of charge, a transceiver may contain software to regulate the operation and data link rates of the communications system, and a fault protection module may monitor various aspects of a spacecraft to ensure proper health and safety of the vehicle.

16.3 Relationship with Spacecraft Hardware

16.3.1 Impact of Hardware on Software

Like typical embedded systems, flight software is tightly coupled with the hardware and interfaces on a spacecraft. For software on the main processor (MP), hardware dependencies fall into three broad categories: (1) on-card, (2) in-chassis, and (3) external interfaces.

16.3.1.1 On-Card

The spacecraft processor is located on an SBC. The SBC can include just the processor or other interfaces such as an SSR or other peripherals. As discussed in [Chap. 15](#), processors for spacecraft are designed to withstand the rigors of the space environment, including high radiation, extreme temperatures, lower power, high g-forces, and vibration of launch. The difficult space environment combined with the long lead-times to space-qualify a processor results in the terrestrial computing community significantly outpacing that of space. For example, whereas a typical desktop operates with clock speeds in the gigahertz range, a spacecraft computer operates in the megahertz range.² These hardware limitations also limit software performance, and require constant assessment of software performance in order to guarantee that the software meets timing constraints. Early in the mission, the system and software engineers should conduct a significant trade or evaluation to ensure that the system's architecture can support the processing, memory, and timing requirements plus an additional margin levied on the processor.

16.3.1.2 In-Chassis Communications

Along with the processor SBC, spacecraft computers can contain multiple other hardware cards to perform

functionality, such as uplink/downlink, spacecraft interface communications, and hardware storage. The MP communicates with these hardware devices via a backplane such as CompactPCI or VMEbus. One typical card is the solid-state recorder that contains banks of memory. It is located either on a local hardware bus or externally to the spacecraft computer. Historically, data recorders were tape drives. This technology has now transitioned to hardware similar to that found in thumb drives. SSRs contain either volatile memory (typically synchronous dynamic random-access memory, SDRAM) or nonvolatile memory (such as Flash or electronically erasable programmable read-only memory, EEPROM). Volatile memory does not retain data when power is removed from the SSR, whereas non-volatile memory does. The flight software interacts with SSRs by storing data either in a file system or in a custom format and then retrieving and transmitting data to the ground upon request. Further information can be found in [Chap. 15](#).

16.3.1.3 External Interfaces

The primary spacecraft computer typically contains only the hardware necessary to perform spacecraft control and data collection. As a result, it requires interfaces to external components. These components can range from instruments to GN&C sensors to communication systems. The hardware interfaces are dependent on spacecraft design constraints such as mass, power, and electromagnetic interference, and are point-to-point interfaces or bus architectures. Point-to-point interfaces include standard serial communications protocols such as RS-232, RS-422, RS-485, and low voltage differential signaling (LVDS). These interfaces provide direct links to components that are not shared. Bus architectures such as MIL-STD-1553B CAN Bus and SpaceWire provide a single input into the MP, with multiple nodes sharing the communications link.

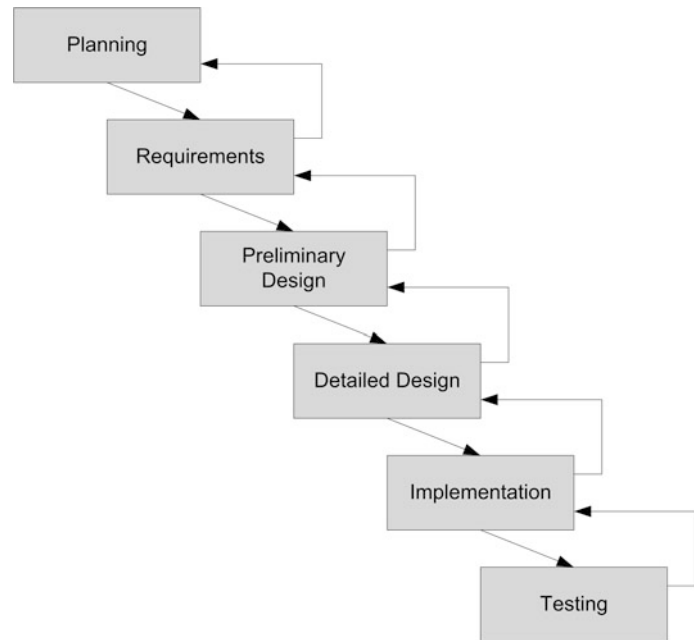
16.4 Flight Software Development Processes

Flight software is a critical subsystem element and must be highly reliable. To achieve this, flight software engineers follow a rigorous software development process to ensure success. Organizations that develop software have unique guidelines, but they all have the same goal of fault-free software and follow the general model described here.³

² Measuring the relative performance of processors cannot be done by strictly comparing clock speeds. Multiple factors such as the instruction set, caching, and floating-point capability can influence performance. Although there is no absolute method for comparison, synthetic benchmark programs such as Dhrystone, Whetstone, and CoreMark can provide acceptable indications of performance.

³ Software development is driven by multiple documents from NASA, the European Space Agency (ESA), and the U.S. Department of Defense (DoD). These include ESA's ECSS-E-40, DoD Standard DOD-STD-2167A, NASA Procedural Requirements 7,150.2, and Goddard Procedural Requirements 1,000.

Fig. 16.1 The waterfall model provides incremental development that feeds back through the process



16.4.1 Development Methodology

The development of software follows an ordered series of activities, or methodology. Although the activities are common to most software development projects, there tends to be variation in order, scope, and level of detail. These deviations are related to the criticality of the software, the amount of information available before beginning development, and the internal processes of the developing organization.

16.4.1.1 Predictive Models

The most common development methodologies in use for spacecraft flight software are predictive models. These models focus on detailed planning. They include Waterfall, Compressed Waterfall, and Incremental Build.

The Waterfall model (Fig. 16.1) describes the development as a single sequence of activities, each building on the preceding activity and culminating in the final product. Some overlap is expected from one activity to the next; however, activities are formally defined, usually occur only once, and culminate in a single product meeting most, if not all, of the original requirements. The Waterfall model is shown as a linear set of steps flowing from one to the other, like a waterfall. In a Waterfall model, each step can impact the previous step, with information attained circling back up the waterfall. For instance, a test case may reveal that a requirement was incorrect, and so that information will be fed back to the requirements phase. Generally, the farther down the waterfall that an error is discovered, the more costly it is to repair.

The Compressed Waterfall model is the same as the Waterfall model except that the detailed design and

implementation steps occur iteratively, beginning with high-risk or highly constrained elements before adding the remaining elements to complete the product. Development teams typically adopt the Compressed Waterfall model when most requirements are known in advance but some risk reduction is required early in the cycle.

The Incremental Build model selects subsets of the software requirements to be delivered as a series of products, or ‘builds’. Software teams use the Incremental Build model when most, but not all of the requirements are known, but there is either a need for early delivery of a partial set of requirements or an unpredictable budget or schedule.

16.4.1.2 Adaptive Models

Adaptive models can also be used in the development of spacecraft software. However, these are usually used only when the software is less critical or on smaller projects. These models focus on adapting quickly to changing program needs. They include Spiral and Agile.

The Spiral model iterates through all of the development activities many times, with each iteration resulting in a new product. Requirements are not completely known at the start of this process and are discovered as part of every iteration. As a result, the requirements and design can be refined each time through the cycle of activities. This model is used when the requirements are vague or ill-defined, and a process is needed to adapt to this.

Agile software development encompasses a number of different methodologies where the software development process must adapt quickly to change and requirements are not well understood at the outset. Most Agile methods are

based on the following tenets: rapid feedback from the customer; a simple, informal process; incremental changes; and small, usually co-located teams. As with the Spiral model, software teams use this more for small and less critical projects as opposed to larger subsystems such as C&DH and GN&C.

16.5 Flight Software Development Steps

Using the Waterfall model as a basis, the following delves into details regarding each step of the development process. Note that documentation does not have a specific section; rather, it is intertwined among the ensuing sections. Each phase has a particular set of documents that should be required in a software development process in order to advance to the next state. All significant documentation—for example, interface control documents (ICD), user's manuals, and requirements documents—requires sign-off by the various interested parties.

16.5.1 Planning

Planning provides the basis for development of the software. During this phase, the software team or lead identifies any areas of potential risk, conducts technology trades, and defines a baseline architecture for the system. In addition, because there is generally a large codebase at the development organization, the software team identifies what can be reused from previous missions. Reuse can reduce development costs and the requirements process, but care should be taken to ensure that the software being reused fits the functionality of the new mission. Also as part of the planning process, the flight software team identifies what tools are going to be used for the various parts of the software effort. This includes scheduling software, documentation and development tools, configuration management (CM) software, and requirements and traceability tools. This type of information is typically saved in a software development plan.

16.5.2 Requirements

Spacecraft flight software requirements typically begin with a flow-down of requirements from the mission system and spacecraft system requirements. Each system-level requirement is reviewed for flight software application. The requirements should include any software safety requirements derived from the system hazard analysis. A unified modeling language (UML) design approach will include use-cases to identify requirements. Additionally, the

concept of operations document for the mission can provide insight into high-level functional requirements for flight software and provide additional use-cases. Once complete, the software development team traces requirements from the mission-level requirement to the flight software requirements.

After the project establishes the high-level software requirements, the software team begins the process of developing detailed requirements. If previous software is available, the developers examine the reused software requirements. Next, the team determines new requirements through analysis. This includes emphasizing functionality, external interfaces, performance, customer/user expectations, and design constraints. The software development team can use ICDs, specifications, and other available subsystem documentation to help understand and refine the requirements.

As part of developing detailed software requirements, the software team creates a context diagram to identify the interfaces and high-level inbound and outbound data flows in relation to the single highest-level aggregate of software. The context diagram consists of a circle representing this single, aggregate software application. Arrows to and from the circle indicate data flow external to the software.

The software engineers derive requirements from the baseline requirements in sufficient detail to perform the architectural design. The requirements should be stated unambiguously; performance requirements should be quantified, and requirements should be written so that each is testable; see [Sect. 7.3](#) for more on this. It is expected that the requirements will be sufficiently thorough that test engineers can understand and evaluate the software built to meet them. However, system designers should take care to not over-specify a system because this may introduce unmanageable overheads.

This activity culminates in the release of a software requirements specification which contains the functional and performance requirements that address what the software must do. It should clearly identify software safety requirements and include any safety-related constraints between the system's hardware and software. The document incorporates the context diagram produced during the scope-determination activity, and it defines the inputs that this software component must process and the outputs that it must produce. It also defines the processing that the software must perform. Often the software requirements specification is evaluated at the software requirements review. After the resolution of any action items from that review, the software requirements specification is baselined and signed off by key members of the project team.

It is also advantageous to involve the test team early in the process. By having testers review requirements with an

eye toward stress testing and testability early, the overall project testing costs can be reduced. In addition, this enables early development of the test scripts and procedures.

16.5.3 Design

After the flight software requirements are baselined, the software development team begins the design process. The design phase encompasses two major steps of the software development process: (1) the definition of the architectural design and (2) the subsequent decomposition into a detailed design.

The architectural design of flight software addresses the structure of the software and the control relationships between components. In addressing these issues, the design must also specify the concurrency of the program, timing considerations for the concurrent components, and the relationships between them. Therefore, the developer of flight software must define the tasks, the communications mechanisms to be used between tasks, and the scheduling of the tasks. Upon completion of the architectural design, the software development team will conduct a preliminary design review (PDR) which typically involves the following materials

- Software context diagram
- Requirements changes since software requirements review
- Software functional overview
- Task communication graph
- High-level driving requirements
- Data flow diagrams
- Text description of the design
- General command list
- General telemetry list
- Processor description including memory capacities and clock speed
- Resource margins
- Risks and mitigation strategy
- The system safety engineer's software safety design analysis
- Developer test approach.

At this point in the software design, the software acceptance test team will often begin reviewing the requirements and planning the software system verification and validation tests.

The final activity in the design phase is the detailed design. This activity further refines the design captured in the architectural design in greater detail. At the end of this activity, the software development team will have a critical design review (CDR) involving the following materials

- Context diagram (defined during requirements specification and updated as needed)
- ICD summary (list of data provided from/to external sources, including commands and telemetry initially produced during requirements specification)
- Changes since PDR
- Task structure charts, flowcharts, or program design language (PDL)
- Software architecture diagram
- Method structure charts, flowcharts, or PDL
- Software interface definition (C language header files or equivalent)
- Design drivers
- Task communication graph (defined during architectural design and updated as needed)
- Test approach (unit, integration, system)
- Estimates of CPU and memory utilization
- Estimates of other hardware utilization
- Risks and approach to mitigate each risk
- The system safety engineer's software safety design analysis
- Test plan outline.

The design products are baselined at the review and updated as part of the response to action items.

16.5.3.1 Meeting the Needs of Reviewers

Given that both the requirements and design phases require the successful completion of reviews, the software development team must be aware of how to meet the needs of reviewers. A successful software review (requirements, design, etc.) should include at least the following

- Review material that conforms to a standard template, delivered with sufficient time for reviewers to prepare
- References and supporting material
- Open issues that are clearly marked as 'to be determined'
- A glossary of terms
- Review agenda with time frames allotted
- Reviewer guidelines for the type of review being conducted.

16.5.3.2 Industry Standards

In addition to processes unique to an organization, flight software development is also governed by industry standards for development.

AS9100 is the quality management standard for aerospace. The Society of Automotive Engineers published the standard in 1999, with input from the American Aerospace Quality Group (AAQG) and support from the International Aerospace Quality Group (IAQG) and the Society of British Aerospace Companies (SBAC). The industry team developed AS9100 to enhance the ISO 9000 series of standards to ensure quality and safety in the high-risk aerospace industry. Specific areas relevant to software that were enhanced include

- Configuration management (CM)
- Design, verification, validation, and testing processes
- Reliability, maintainability, and safety
- Product documentation
- Corrective action
- Inspection and testing procedures.

Capability maturity model integration (CMMI) is a set of best practices and models that are used to improve an existing process. Most software development organizations use CMMI to take an existing process, such as AS9100 or ECSS, and improve it by adding practices such as collecting and managing requirements, measuring performance, planning work, and assessing risks.

16.5.4 Implementation

16.5.4.1 Languages

As stated previously, on early spacecraft, engineers relied on either assembly or custom languages to develop software. As the underlying hardware has progressed to more mainstream architectures, the need for custom languages has diminished. Hence, software teams predominantly use higher-level languages such as C, C++, or Ada. Using these types of well-known languages greatly increases the potential developer base, expands on the wealth of previously developed software components and development tools, and enhances the readability and reuse of code as well as overall productivity. Software engineers do still resort to assembly language for low-level drivers or when necessary to increase the speed of execution for certain aspects of the code.

16.5.4.2 Automatic Code Generation

As spacecraft on-board processing resources have increased, the need to optimize code has decreased. As a result, code for GN&C software is increasingly being generated automatically from the high-level system specification produced in tools like MATLAB[®] Simulink and dSPACE TargetLink. Such techniques have been used on low-cost missions such as LISA Pathfinder and Proba in order to reduce the time necessary to redevelop the software in response to late changes in the specification. However, use of this technique can make software verification more complex because the software engineers must understand code that they have not written.

16.5.4.3 Development Environment/Operating Systems

Most modern spacecraft use a real-time operating system (RTOS). Responsibilities of the RTOS include managing the various tasks in a system, handling operating system constructs such as semaphores and queues, and interfacing

with the underlying hardware. A key feature of an RTOS versus a non-RTOS is that it enables scheduling and execution of tasks with a minimal amount of jitter. An RTOS also provides a well-documented and understood method for handling errors and faults. Although the majority of an RTOS can be used from platform to platform, an RTOS uses an additional hardware-specific set of software called the board support package (BSP) to adapt to a new hardware domain. A BSP is specific to a particular platform; therefore, it is not reused outside of the specific hardware.

16.5.4.4 Time and Space Partitioning of Software Systems

As software systems become more complex, the integration and validation of components of different criticality and with different origins becomes ever more expensive. To address this, the space industry is starting to adopt a technique called integrated modular architecture (IMA) that is used in the commercial aerospace industry. The technique relies on software/hardware support to provide a number of partitions that guarantee isolation of a set of components from each other. This is to prevent one software component from corrupting the memory or stealing computer resources used by another component. This simplifies development and validation because each partition can be developed in isolation. The use of time and space partitioning to divide a typical software system is shown in Fig. 16.2.

16.5.4.5 Specific Flight Software/Embedded Techniques

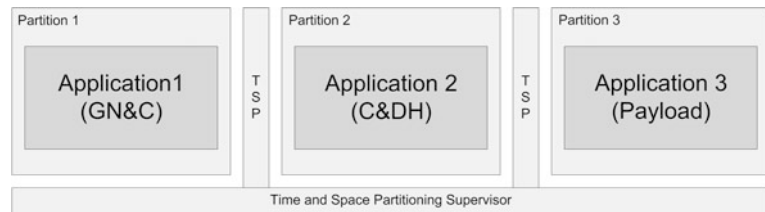
Embedded software, and spacecraft software in particular, requires programming techniques that are different from those used in a workstation environment. This is due both to the high-reliability requirements of flight software and to the constrained environment seen on the spacecraft. As such, developers follow stringent coding guidelines either created within the organization or externally, such as that developed by the Motor Industry Software Reliability Association called the MISRA C coding guidelines. Typical recommendations include avoiding dynamic memory allocations, limiting the use of global variables, and refraining from recursive functions.

16.5.4.6 Configuration Management

Configuration management (CM) is a method by which software code is stored in a common repository. Developers check out this code to make changes and then commit the code back into the repository. A CM system enables features such as safe storage of the software code, versioning of the software, and revision history. Several CM systems are available commercially and open-source.

Only the smallest of software tasks should be performed without CM, so it follows that flight software is highly dependent on CM. This is not only for switching between

Fig. 16.2 Partitioning software systems



versions during development but also for delivering software to the spacecraft. The software development team must know which versions are used to run each set of tests and which are loaded and delivered to mission operations. A valuable addition to this tracking is to include an element in the software that can display date, time, and versioning information in telemetry.

16.5.4.7 Static Analysis

In addition to human review of software, flight software is subjected to static analysis using various tools. These can be particularly helpful in automating code reviews. Instead of focusing on mundane checking of logic (such as inclusion of breaks in a switch statement), reviewers can instead study more in depth the functionality of the code.

16.6 Flight Software Development Testing

16.6.1 Testing Philosophy and Approach

Given the critical nature of flight software, testing has a high profile in the overall development process. This testing is performed not only by the developers but often also by an independent team that answers to an authority outside of the normal program management structure. The tests on both sides of this organizational divide are generally defined and created by referring back to a well-thought-out set of operational and performance requirements. The tests need to be comprehensive, traceable back to the requirements, auditable, and repeatable.

16.6.2 Test Simulators

One of the most fundamental aspects of flight software testing is the need to perform the testing and validation in an environment that is as representative as possible of the actual spacecraft. In the absence of the real spacecraft, this is often achieved by testing within a so-called hardware-in-the-loop (HIL) configuration that provides realistic models of real-world sensors and actuators as well as the corresponding mathematical models of the spacecraft dynamics and environment. A high-fidelity HIL test bed is generally required; however, as necessary as this might be, it is also

expensive. The reality of modern-day budgets often means that there are only one or two HIL environments available for the entire spacecraft testing effort. The traditional method used to mitigate the HIL bottleneck is to develop and deploy so-called software-in-the-loop (SIL) simulators with which the sensors and actuators are mathematically modeled along with the spacecraft dynamics and environment. In a SIL environment, testers can develop scripts and prove their test methods before inevitably executing and verifying them, once again, in the presumably much higher fidelity of the HIL environment. A SIL is normally implemented in off-the-shelf desktop computers and is therefore inexpensive from a hardware cost perspective. It can be expensive from a software perspective because the software must be carefully designed and tested before being released for validation of software of similar complexity.

16.6.3 Testing Steps

The software and test team perform testing throughout the development process.

16.6.3.1 Unit

Individual developers perform unit testing on their specific software units. This is done at the procedure/function level and includes verification of all paths in the code. Software engineers conduct this testing either through manual coding or via automated testing software.

16.6.3.2 Application Test

As with unit tests, individual developers conduct application testing. Although similar to unit testing, these tests do not focus on specific path testing but instead examine the full functionality of an application. Developers are verifying that the application operates as a full unit and meets the unit requirements.

16.6.3.3 Integration Test

Upon delivery of the individual applications, the integration tester, who can be an individual such as the software lead or the entire team, conducts tests over the complete system. This again builds on the incremental nature of testing—from unit to application to system. The integration testing consists of tracing and verifying the requirements of the system as a whole.

16.6.3.4 Acceptance Test

The acceptance test process has the same goal as integration testing: tracing requirements to tests. However, the acceptance test team is independent of the software team. Hence, the tester's ability to discover either bugs or requirement discrepancies is enhanced through a tester seeing the system from a different viewpoint. Acceptance tests can be either scripted, or hand operated. In both cases, testers typically use either a modified or a fully implemented version of the ground system to operate the system through commanding and viewing telemetry.

16.6.3.5 Scenario Test

Scenario testing can be viewed as either a complement to acceptance testing or an alternative. With scenario testing, a tester may not map specific requirements to a specific test but instead focus on either groups of requirements or use-cases of system functionality. For example, instead of verifying that a specific command increases a specific counter, a scenario test would involve multiple operations, of which that specific command is just a small part.

16.6.3.6 Stress Test

As with scenario testing, a stress test focuses on the larger system. Instead of verification, however, a stress test attempts to break the system by adding high loads, where possible. This includes activities such as triggering instruments to send their maximum amount of data, executing the full complement of fault protection algorithms, and recording large amounts of data to the SSR—all at the same time. This stresses the system to determine whether there are any failures in timing, operation, or functionality.

16.6.3.7 Operational Testing

After delivery, the flight software team contributes to additional operational testing. These tests include launch simulations, day-in-the-life testing, and off-nominal testing.

16.7 Post-Launch

16.7.1 Software Updates

Modern spacecraft rarely exclude the ability for the mission operations team to update the software post-launch. As such, the flight software team needs to include the ability to modify the software during the actual mission. Software updates range across the following from least to most intrusive

- In-line patch
- Jump patch
- Image load.

To conduct an in-line patch, the software team pokes a location in memory with either a new instruction or a modification to a data element, thus changing the operation of the software. Patches that are more complex may entail a jump patch. With this method, the team modifies the executing code as it operates out of random-access memory (RAM) by loading new code into areas of unused RAM and then using assembly language modifications to jump to that location. The disadvantage of both of these techniques is that they usually require lower-level programming skills, the modification is harder to validate, and the patch is lost each time the processor is reset and the stored image is reloaded. An alternative to patching is to load a new image into a non-volatile memory location while running the older software version out of static random-access memory (SRAM). Once the success of the upload has been verified, the operations team sends a command to tell the computer to boot from the new image and then resets the spacecraft computer. Although this provides for consistency in code images and uses the usual tools that made the original image, resetting the spacecraft can be difficult and, for some missions, not cost-effective. Intentionally resetting the spacecraft usually requires heavy mission operations involvement and a significant period to transition back to operational mode.

16.7.2 Handling Post-Launch Anomalies

The most important element in handling post-launch anomalies is for the software team to ensure that it has the data needed to debug issues in flight. This capability begins during design by including an anomaly or event-logging process within the software architecture. The event logging includes sufficient information to make the anomaly unique and identifiable to a particular application. Also, the software architecture should include the ability to store the most recent anomalies in a memory location that the mission operations team can retrieve after a reset. In addition to the event log, the software should store information such as the reset cause, the task switch log, and processor stack information.

16.8 Spacecraft Software Architectures

A spacecraft software architecture defines how the various software elements within a system operate and communicate. The architecture is a product of the preliminary and critical design phases of the software development process and is essential to producing software that is efficient to develop, testable, and maintainable. The software team documents an architecture beginning with an overall top-level design and continues with each sub-element further

delineated by more detailed designs. The top-level design shows how the software is grouped into an overall computer software configuration item (CSCI). The sub-elements are the computer software components (CSC).

As defined by IEEE's Standard Glossary of Software Engineering Terminology, a CSCI is "*an aggregation of software that is designated for configuration management and treated as a single entity in the configuration management process*" [3]. The flight software architect defines the breadth of a CSCI, and this varies given organizational or functional constraints. A CSCI can be broad and typically consists of the software binary code that is loaded into a given processor on the spacecraft. Depending on the overall architecture, that can include boot, C&DH and GN&C, if operating on the same processor, C&DH and GN&C alone if each is operating on separate processors, and the various software packages that operate on instruments or other subsystem elements. The CSCI can also be more tightly defined by decomposing the larger subsystem into smaller theme-based CSCIs such as commanding, telemetry, and instrument processing. Within each of these CSCIs are applications that perform related functionality.

A CSC is defined as "*a functionally or logically distinct part of a computer software configuration item, typically an aggregate of two or more software units.*" A flight software architect deconstructs the requirements related to the CSCI into a number of CSCs, each of which contains one or more tasks to fulfill the corresponding requirements. Defining distinct CSCs enables the lead software engineer to cleanly assign responsibilities for the various modules across the team.

Once the software architect delineates the CSCs, the next step is to determine a method of communication between elements. This architecture construction generally falls into two categories: (1) tightly coupled and (2) decoupled software. In a tightly coupled system, there are specific communications mechanisms between the various software elements. These include direct function calls into their respective CSCs' code space, message queues contained in the RTOS, or shared memory spaces. Although this is a straightforward solution, a disadvantage of a tightly coupled architecture is that changes can affect multiple CSCs and reuse can be difficult because the code is heavily linked together. With decoupled software, the CSCs stand apart from one another, with a single direct line of communication in and out of the system. A standard architecture for a decoupled system is based on message passing. In this architecture, a software bus routes messages between CSCs. The advantage of this architecture is that each CSC can stand alone, and revised CSCs can be dropped in and out of an architecture as long as they conform to the appropriate message interface.

An example of a decoupled architecture is shown in Fig. 16.3 and represents most of the main capabilities seen in the main computer of a spacecraft, with each bubble

representing a CSC. 'Scheduler' is responsible for initiating the various tasks at predetermined rates. The rates typically range from once per second to tens of times per second. 'Uplink' ingests commands and other data from the communications system and then distributes those data throughout the system. 'Command manager' controls the command rate of the system and determines the priority of commands. 'Autonomy' watches over the spacecraft operation and health and safety and acts when necessary to recover from faults or, if necessary, to save the spacecraft. 'Time tags' keep track of spacecraft MET and execute a series of commands when the MET reaches a corresponding value. 'Instrument manager' controls the flow of commands to various instruments and collects science data from these sources. 'Spacecraft interfaces' are responsible for collecting sensor data and other non-instrument data from outside of the main computer. 'GN&C' controls the attitude, thrusters, and other dynamic positional characteristics of the vehicle. 'Record' stores both spacecraft and instrument housekeeping data to the SSR. 'File manager' controls file system operations on the SSR. 'SSR playback' retrieves data from the SSR for transmission to the ground system. 'Downlink' manages the output of telemetry from the spacecraft. 'Memory scrub' continually reads data from the memories of various computers to enable data correction. 'CPU performance monitor' keeps track of how much loading the CPU sees on a 1 s basis. The CSCs are layered on multiple interfaces, including a software message bus, the operating system, the processor hardware, and the hardware communication bus.

16.8.1 Boot

Boot enables operation of the single board computer and peripherals. The boot software is generally created earlier in the development cycle since all later applications rely on boot to provide an operating platform. The boot software can reside on a 'write once' device such as a programmable read-only memory (PROM) or it can be included a rewritable non-volatile memory location. Since all operation depends on boot, systems developers should take care to minimize complexity in the boot process and reduce boot to only the essentials needed to enable operation of the upper layer programs.

16.8.2 Command and Data Handling

Nearly all of the applications in the example architecture in Fig. 16.3 fall into the category of the C&DH subsystem. The primary role of the C&DH subsystem is to ingest commands from the radio frequency (RF) link, record housekeeping and science data, and to return those data in

the form of telemetry to the ground. Generally, C&DH is also a ‘catch all’ for other functionality, including autonomy, mission planning, data compression, and image processing—essentially any element of spacecraft software that does not directly involve GN&C.

16.8.2.1 Commanding

Multiple sources can generate commands for a C&DH system, including operators on the ground, other spacecraft, other processors, and stored commands. For commands received through the RF link, command input is relatively slow. On a typical robotic spacecraft, the command rates are low, typically ranging up to only 2,000 bps. Hence, a command ingest application does not require a significant amount of processing time, and the interrupt rates for commanding to receive new data are relatively low. Commands received via this route are usually encased in protocols defined by the Consultative Committee for Space Data Systems (CCSDS). These protocols frame and encapsulate the data in multiple layers to ensure proper transmission and receipt. Two types of methods are used throughout spacecraft design: (1) telecommands (TC) and (2) advanced orbiting service (AOS) [4]. Although older than the AOS protocols, TC is still widely in use today. Parsing the CCSDS data as they come into the system is performed by either hardware or software. If performed by software, the responsible C&DH module peels back the protocol layers until it extracts the commands.

Command constraints are a critical element in designing a C&DH system, and they can include

- *Ordering*—some commands are order dependent. For instance, the spacecraft may need to slew before an imager takes a photo.
- *Command execution time*—commands vary in execution time. For example, a command that simply sets a 1-bit flag will execute quickly, whereas, a file system operation that may move megabytes of data will take a significantly longer time to operate. The flight software design must accommodate the nondeterministic nature of commanding.
- *Failure processing*—commands may fail for multiple reasons. For example, the command may have been corrupted in non-volatile memory, the state of the system may not have been correct when the command was executed, or the command may have had incorrect arguments. Commands are evaluated on an individual basis to determine whether a command failure is critical.
- *Command prioritization*—some commands take priority over others. Real-time commands from the ground are the highest priority because they allow for preemption of any commands that may be producing critical errors.
- *Time criticality*—although some instruments may not require precision timing for activation, a low-orbiting

high-resolution imager may miss a target without millisecond precision in activating the camera. Hence, whether a command can execute within a given second or at a certain millisecond is dependent on the instruments or configuration of the spacecraft.

Commands can come in the form of multiple sources and for multiple purposes. Real-time commands are sent by the operations team via the ground system. The flight software gives precedence to this type of command over others to allow for direct communication by the ground team. This allows mission operations to override any stored commands that may be in the midst of executing, and it provides a mechanism for immediate interaction with the flight software. A spacecraft also may have a set of time-triggered commands. These are one or more commands that execute when the flight software recognizes either that a given time has elapsed or that a certain time has been reached. Time-tagged commands are especially important for deep-space missions because of the frequent communications outages with ground operators and long light-trip times. Time-tagged commands are similar in nature to the pre-stored sequencer commands in early spacecraft missions.

Most spacecraft have some form of autonomous behavior to monitor spacecraft health and ensure that an appropriate response is executed in the case of any failures or unexpected events. Autonomy can range from rudimentary rule checkers that monitor and compare telemetry against a pre-stored set of values to complex artificial intelligence systems that rely on technology such as expert systems.

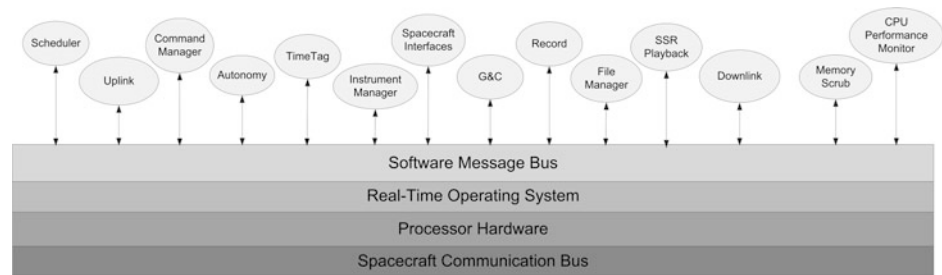
A macro is any functionality that allows large sets of commands to be stored on-board the spacecraft and then triggered by a single telecommand. They can also be triggered by other elements of the system, such as time tags, autonomy, or real-time events. The use of macros can significantly reduce the amount of communications bandwidth required to execute in real time because they are already stored on the spacecraft before executing.

Scripting allows for more complex on-board operation than can be attained through simple macros. A script can take a form similar to programming languages and include constructs such as ‘while’ loops and ‘if’ statements that must be interpreted on-board. However, although it increases flexibility, a scripting language can require more processing time to interpret the symbols of the script and must be prevented from executing typical programming errors such as variables overflowing.

16.8.2.2 Telemetry

Telemetry is the data returned from the spacecraft. Telemetry consists of housekeeping data and science data. Housekeeping contains information concerning the health and safety of the spacecraft, instruments, and payloads. A key element in designing the flight software telemetry is to

Fig. 16.3 A layered architecture enables separation of functionality



determine the frequency and priority of the various elements. Priority in the housekeeping data should be given to highly dynamic data as opposed to data that typically are unchanged unless by command; this reduces the overhead of sending data that rarely change. The flight software collects these data in the form of packets and frames. As with commanding, most spacecraft use CCSDS standards as the basis for this process. AOS is again an option, as is the CCSDS telemetry (TM) standard [5].

TM has less layering than TC because it consists only of a frame header and a frame footer. Framing can be performed via either hardware or software. If software is used, a telemetry application collects formatted data from the various subsystems and tasks and encases the data in TM frames. For this example, data are contained in variable-sized CCSDS packets. Variable- and fixed-sized CCSDS packets have their respective advantages. For example, variable-sized packets can reduce bandwidth waste because packets are sized to only the data needed, but fixed-length packets can ease ground processing and frame creation.

As robotic spacecraft are rarely in constant communication with the ground system, the flight software must record telemetry that will then be played back when the RF system allows. Modern spacecraft use an SSR for this purpose. At its core, the SSR consists of a large array of either volatile or non-volatile memory. An SSR can range from several megabytes to hundreds of gigabytes. Depending on the location of the SSR, the flight software maps the memory of the SSR for access by the recording task. The flight software records data either by storing the data raw in customized data structures or through the use of a file system. Multiple factors affect the architecture for storing data. Raw partitions typically use less RAM on the on-board processor and require less complex software to manage. A file system gives more flexibility because it uses common interface calls for opening, reading, writing, and closing files. However, file systems tend to require more RAM to store the structure of the file system. The recording application must also accommodate the type of memory used to store the data on the SSR. Two types of memory are primarily in use on SSRs: (1) SDRAM and (2) Flash. SDRAM requires less complex software because it can be

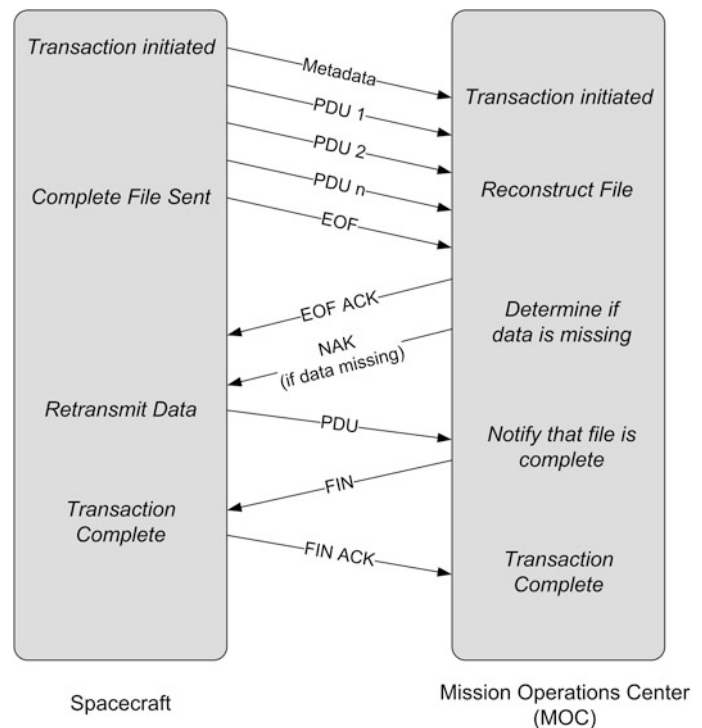
read and written similarly to the SRAM on a flight processor, albeit with slower access times. However, SDRAM is volatile, and the data are lost when power is removed from the SSR. Flash, on the other hand, is non-volatile but requires more complex software to access the data. When writing to Flash, an entire sector (the minimum unit of erasure which can be multiple bytes) must be erased before storing the data to memory, as opposed to modifying single bytes. Additionally, the software must accommodate wear leveling because Flash will lose functionality if written too often.

Once a downlink session is available, the mission operations team instructs the flight software to downlink the data. The software application for downlinking the data depends on how the data have been stored. If the data are stored as raw packets, the software can retrieve the data from the SSR and insert the packets directly into the telemetry stream. When using CCSDS TM, playback data are placed in a different virtual channel to allow the ground software to distinguish between real-time and playback frames. If the data are stored as files, the flight software can use a mechanism such as the CCSDS file delivery protocol (CFDP), a protocol similar to the file transfer protocol (FTP), which enables transmission of files between the ground and spacecraft systems [6]. A typical CFDP transaction is shown in Fig. 16.4.

To process a CFDP transaction, the flight software breaks a file into multiple pieces called protocol data units (PDU), which are framed and sent to the ground one by one in the form of file information and file segments. Once the file has been sent, the spacecraft notifies the ground that the transmission is complete by sending an end-of-file (EOF) marker. The ground acknowledges receipt of this information via an acknowledgment (ACK) PDU and then retransmits missing pieces via a negative acknowledgment (NAK) or, if all data have been received correctly, instructs the spacecraft to terminate the transaction through a finished indicator (FIN). The spacecraft then responds accordingly by either resending data or closing the transaction.

In designing the playback system, the flight software architecture must account for the additional processing time required to extract data from the SSR as well as maintaining

Fig. 16.4 CFDP enables FTP-like downlink of files



pace with the downlink rate. Because the SSR is usually located off of the processor card, SSR accesses are slower than direct RAM access. Unlike the low rates for command uplink, the playback task must accommodate higher rates from hundreds of kilobits per second to several megabits per second depending on the spacecraft capabilities. This creates a high interrupt rate as the RF is requesting frames more frequently and introduces a large amount of processor overhead in creating the frames. At sufficiently high downlink rates, system developers may target a trade that shares the downlink processing between software and hardware.

16.8.3 Communication Interfaces

To collect data from instruments or transmit and receive data from other subsystems, the spacecraft flight software communicates across various electrical interfaces. These interfaces include dedicated serial links as well common buses, such as MIL-STD-1553B or SpaceWire which are discussed in detail in [Chap. 15](#), and summarized here from a software perspective.

16.8.3.1 Serial Links

Spacecraft use serial links similar to those in terrestrial applications. Spacecraft designers use three primary protocols for serial communication: RS-232, RS-422, and RS-485. RS-232 is the oldest of these protocols and the simplest. Until recently, RS-232 ports were also seen frequently

on personal computers. RS-422 and RS-485 are also heavily used for industrial platforms.

16.8.3.2 Mil-std-1553b

The MIL-STD-1553B bus interface originated in military avionics. It has a redundant interface with time-division multiplexing and the ability to communicate with nodes on a network. Activity on a 1553 bus is controlled by the bus controller (BC), typically on the main processor of the spacecraft. The other nodes on the 1553 network are called remote terminals (RT). The BC instructs the RTs either to receive or transmit data as needed. The BC selects one of the data lines on the bus to be the prime, and when a transaction fails on that bus, it is retried on the different wire. A bus monitor can be included in a 1553 architecture to observe bus activity.

There are multiple types of transactions on a 1553 bus. These include instructions to transmit and receive data, broadcast data from the BC, mode code commands, and RT-to-RT transfers.

A 1553 bus can be either synchronous or asynchronous, although most systems are synchronous. A key design element for a synchronous 1553 design is the bus schedule, which consists of a major frame that occurs at 1 s intervals and multiple minor frames that occur at a sub-second interval such as 50 Hz. During each minor frame, the BC executes a set of transactions that are performed during that sub-second. The 1553 software designer must design the bus schedule to work within the constraints of the 1553

bandwidth to ensure that all transmissions can be accommodated during the interval and also allow for margins in the event that the bus schedule is modified after launch.

Because 1553 has been in use for a number of years, there are multiple chip manufacturers that have developed processors that perform the lower-level protocol operations. This alleviates much of the work required by the flight software.

16.8.3.3 SpaceWire

SpaceWire is similar to 1553 in that it provides a common interface to the various elements on a spacecraft network, but it has a much higher bandwidth and does not require the time division of the 1553 bus. The advantage of the latter difference is that data can be sent when available, but it can also make operation of the bus less deterministic and prone to collisions. This requires care in planning bus operations to prevent, for example, a large but low-priority instrument packet that blocks delivery of a more critical packet such as a thruster command. The flight software can alleviate some of this by introducing additional processing to ensure that transactions do not conflict—essentially introducing a bus schedule capability of time division. However, this can reduce bandwidth given that the time is reserved regardless of whether or not a transaction occurs and can introduce jitter and delay as a transaction awaits an opportunity to transmit.

Collisions on the SpaceWire bus can also be avoided through the use of alternate (redundant) paths dedicated to high-priority packets and/or ensuring that the paths of these packets do not overlap. Group-adaptive routing is a method for automatically switching packet delivery to an alternate path if the primary route is unavailable because of congestion or link failure.

Another important difference is that the 1553 specification mandates redundancy through a pair of links and automatic error detection and switching between them. SpaceWire requires explicit planning to ensure the physical availability of alternate paths and, with the exception of group-adaptive routing, software protocols to ensure reliability. Several protocols have been proposed that include error detection and packet transmission; however, these protocols have not been fully standardized at this time and, unlike 1553, are not an inherent capability of the system.

16.8.4 Additional On-board Processing

As the processing capability of processors increases, additional applications are being added to the spacecraft computer that extend beyond typical data-handling and guidance and control tasks. Note that the GN&C application is considered separately in [Sect. 16.9](#).

16.8.4.1 Compression

Downlink bandwidth is scarce, particularly for deep-space missions, and the greatest use of this bandwidth is science data. The flight software reduces this strain on the system by compressing the data. Data compression comes in two forms: (1) lossless and (2) lossy. Lossless data compression is less efficient than lossy compression, but, upon decompression, it retains the dynamic range of all of the data recorded. Lossy compression can reduce the amount of data played back, but at the cost of losing some of the data detail. The use of lossy or lossless data is data dependent. For instance, a lossy image may result in an acceptable loss of resolution, whereas lossy data for another instrument may make the data useless. In both cases, loss of some of the data in transmission can result in uncorrectable data on the ground. When compression is used, reliable transmission of data through capabilities such as CFDP is particularly important.

16.8.4.2 Image Processing

An alternative to compression is performing on-board image processing to reduce science data and download only the scientifically interesting data. For example, an experiment on the Mars Odyssey spacecraft took images of the Mars polar ice cap to determine its size and location. Instead of downlinking every image, the on-board software calculated the latitude of the ice cap and only transmitted that information. It is also possible to develop software to distinguish characteristics of rocks such as structure, color, and texture, and use these characteristics to identify images that contain features that are unusual and that may interest the scientists on the ground.

16.8.4.3 On-board Mission Planning

Along with autonomy, another area where flight software developers are applying the capabilities of artificial intelligence is mission planning. To make operations simpler and more flexible, on-board planning software enables the spacecraft to be directed to perform high-level tasks rather than sending large numbers of very detailed commands. The software is then able to autonomously plan the observations on the basis of this information and knowledge of the spacecraft's current status. This enables the spacecraft to quickly adapt to changing conditions by replanning the mission operations if circumstances change or to observe scientifically interesting events that could not be identified with the delay in ground communications.

16.8.5 Other Subsystem Support

Although not as extensive as the primary computer architecture, software is also prevalent in additional subsystems

besides C&DH (Sect. 16.8.2) and GN&C (Sect. 16.9). Instances of software outside of the main process include

- *Power*—a power system may need to perform peak-power tracking and battery performance analysis.
- *RF*—engineers are increasingly turning to software to further expand communications capabilities (in particular, software-defined radios that can be reconfigured as needed to suit a mission’s needs).
- *Instruments*—some instrument software is reaching the complexity of the main computer, to include compression software, data recording on their own SSRs, and image processing. At a minimum, instruments must have the ability to execute commands, return telemetry, and interface to the main computer.

16.9 Guidance, Navigation, and Control

GN&C is a control system. The primary functions of its subsystems are to maintain spacecraft attitude, to execute propulsive maneuvers for spacecraft trajectory control, and to provide a navigation function that maintains positional knowledge within a given frame of reference. The GN&C subsystem is discussed in detail in Chap. 12.

The GN&C software accumulates data from various devices to control the spacecraft. For example, sensor inputs such as IMUs and STs provide body rate information and knowledge of the inertial orientation of the vehicle. Actuator outputs such as thrusters, torque rods, and reaction wheels (RW) exert forces on the vehicle to establish the desired attitude or, in the case of a propulsive maneuver, to change the vehicle’s velocity along a given vector in inertial space.

Navigation refers to the determination, at a given time, of a vehicle’s position, velocity, and attitude, commonly referred to as the vehicle’s ‘state vector’ or, simply, ‘state’; see Sect. 4.1.4. Guidance refers to the ability to determine the change in the vehicle’s state to take it from the current state to a commanded or target state. Finally, the control task directly commands the actuators that physically move or rotate the vehicle on the basis of inputs from the sensors as well as commands from the guidance task. Typical GN&C sensors include aforementioned IMUs, STs, and SSs. Typical actuators include RWs, magnetotorquer rods, and attitude control thrusters as propulsion.

Flight software development for GN&C generally involves both flight software developers as well as GN&C analysts who have a high level of knowledge of spacecraft dynamics. As a result, the GN&C flight software includes algorithms developed with input from the GN&C analysts plus the ‘wrapper code’ that enables integration into the overall flight software architecture. In general, the GN&C algorithms are provided to the flight software team in one of

two ways: either in the form of an algorithm description document (ADD), from which the algorithms are directly converted to flight software; or more typically, in the form of a pretested software library or set of functions that have been auto-generated from a development environment used by the analysts. The wrapper code, then, is typically responsible for integrating the auto-generated GN&C algorithms, acquiring and formatting time-stamped sensor data, executing the algorithms, and formatting and applying the corresponding commands to the actuators. It is also responsible for forwarding commands from the C&DH and providing often extensive telemetry back to the C&DH for downlink to the ground.

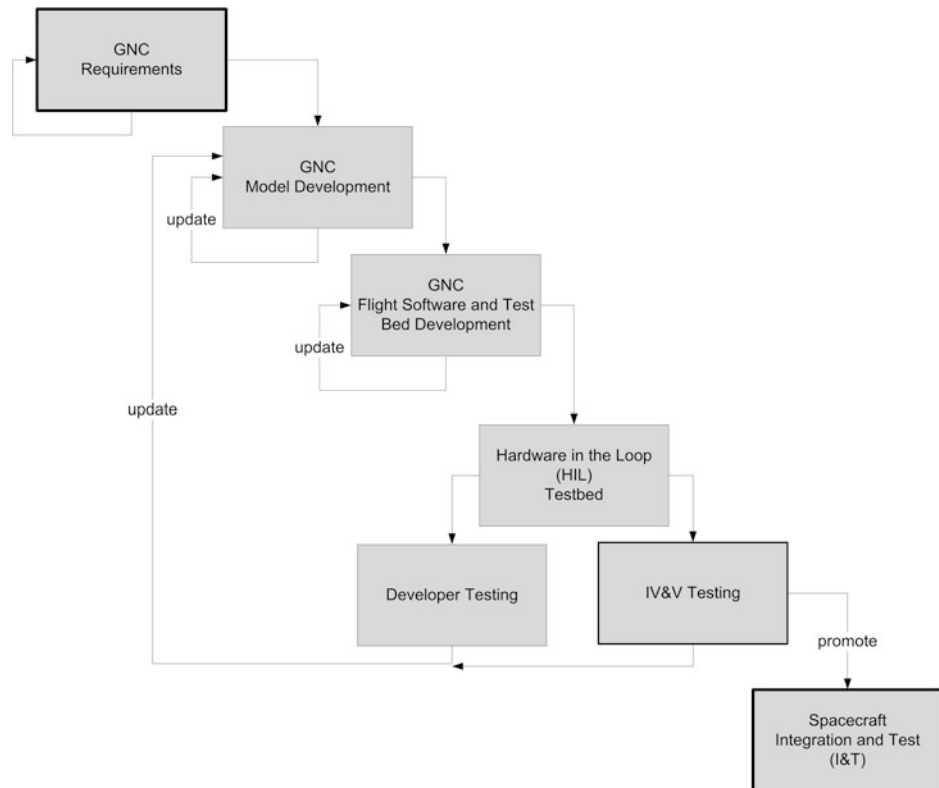
The software executes the GN&C algorithms at a fixed rate and with typical execution cycles of 50 Hz for control and 1 Hz for guidance and navigation. Critically, the internal clock from which the execution rates are derived is, in turn, synchronized to MET, UTC, or some other time standard.

From a GN&C perspective, the architectural impact of a software bus system is that it is data-driven, whereas the GN&C, as stated above, can be perceived as a synchronous control loop. It is up to the software designer, as well as the GN&C analysts, to ensure that these two very different approaches do not conflict. In other words, it is important that sensor data, even though they are time-tagged, be read and processed in such a way that they are provided to the control and guidance tasks within the time boundaries expected for those data. Similarly, it is important that any output control data to RWs and attitude control thrusters not be delayed.

16.9.1 Process Flow

Although there is overlap between writing code for C&DH software and GN&C software, the GN&C process flow is covered here to illustrate the unique aspects of this system’s development. Figure 16.5 shows a high-level version of the process that leads from the initial definition of requirements to the final delivery of the GN&C software to the spacecraft. Each part of the process can be viewed as iterative. The GN&C analysts develop the algorithms from the requirements while continually testing within their workstation-based environment until a working version can be passed to the flight software and test bed development areas. Those areas, in turn, integrate the algorithms and develop the associated wrapper code for acquiring sensor data, reading and applying actuator data, commands, and telemetry. Once unit testing is complete, the combined algorithms and wrapper code in both the flight and test bed are integrated into a HIL test environment or, alternatively, into an intermediate SIL test environment. Developers now have a complete test

Fig. 16.5 GN&C software development process flow with independent verification and validation (IV&V) testing



environment that includes sensor and actuator models as well as the capability to send commands and output telemetry. In parallel to the ongoing development effort and unit tests, independent testing of the algorithms can now also proceed in such a way that as the developers continue with unit testing of one software build, independent testing is validating the previous build. Finally, once all of the functionality has been included and independently tested, the GN&C software is integrated onto the spacecraft as part of the spacecraft integration and testing (I&T) activities.

16.9.1.1 Software Development of GN&C Algorithms

There are, in general, two main approaches commonly in use for the development of GN&C algorithms. The first and less common approach is that the algorithms are functionally and mathematically described by GN&C analysts in the form of an ADD from which GN&C software specialists convert the algorithms into flight code. The ADD might include example code that has been used to validate the algorithms within the analyst's development and test environment. The advantage of this approach is that the GN&C analysts can focus on the specifics of the algorithms and not then be encumbered by the process associated with the development of mission-critical flight code. The disadvantage is that the software developers themselves must have sufficient skill to understand and interpret the algorithm in order to properly design, code, and

then unit test the algorithms, which is a skill set that is not often available in the workforce.

The second and most common approach is one in which the GN&C analysts develop the algorithms within a model from which the algorithms are subsequently auto-generated into C or C++ code. This code is then provided as a library or a self-contained set of functions and methods that encapsulate the entire set of GN&C algorithms. The tool set of choice is typically a commercial product within which an analyst can design, simulate, and test the GN&C algorithms and models. The main advantage of this approach is that the algorithms as developed and tested in the modeling environment are incorporated directly into the flight code without interpretation by a software developer. The flight software engineer instead focuses on developing the wrapper code, which integrates the model into the overall flight software by acquiring and providing inputs to the models, executing the models, and then reading the corresponding outputs. The disadvantage of this approach is primarily one of visibility—the auto-generated code is typically not readily readable or accessible. Instead, it becomes a 'black box' that accepts inputs and generates outputs, but the system depends on the reliability of the auto-generated code. Hence, the confidence in the design from a software perspective comes down to testing and verification that the performance of the models is identical to that of the workstation environment that was used by the analysts.

16.9.1.2 Model Construction

In using either technique for model development, the resultant software must be constructed within the confines of an embedded environment. The models cannot be a loose and haphazard collation of algorithms but must be a well-designed architecture within which inputs, parameters, and outputs and model rates are clearly defined and consistent with the flight environment. It is a primary assumption that the GN&C analysts work closely with the software developers to structure their models to target an embedded system and to enforce naming conventions on inputs and outputs as well as internal parameters that will support external auto-coding methods in support of the auto-generation of flight and ground databases.

16.9.2 Implementation Issues

16.9.2.1 Time

As stated above, because GN&C is a control system, the execution of the GN&C algorithms is tightly controlled, meaning time-tagging of input data and the corresponding model execution must be synchronized to a known time epoch, generally provided by an internal clock with the flight computer. The time tag can be in many formats, most commonly MET, which counts in integer seconds from a known epoch; others include the CCSDS unsegmented time code (CUC), International Atomic Time, and Terrestrial Dynamical Time. Although there are many choices, the fundamental basis of all of them is that they provide a contiguous measurement of time referenced to a known time epoch, such as the J2000 epoch (see Sect. 4.1.6), with some offset that may have to be periodically adjusted. Time and the measurement and management of it on a spacecraft is a complex subject. However, purely from a software perspective, the goal is to provide the GN&C subsystem with time-tagged sensor inputs, execute the models at a known rate consistent with that epoch or at least with knowledge of that epoch, and then apply the corresponding actuator outputs at the designated control rate.

16.9.2.2 Model Execution

GN&C models must execute at the rates defined by the analysts. This can be anywhere from 1 up to 100 Hz but is typically 50 Hz. In a 'simple' GN&C system, inputs are 'gathered' at every model 'step' (e.g., every 20 ms for a 50-Hz control rate), the model is executed, the model takes some time less than the 20-ms period allocated to it, and, on return, the outputs are gathered and applied to the respective actuators, and/or output as telemetry. The priority of the single task must be set high enough within the operating system to guarantee that the control task completes within

the 20-ms period or else a system-level anomaly is raised that may lead, in some cases, to a processor reset. This typically means that the GN&C task priority is set to second or third highest right after device drivers.

In a more complex GN&C system that includes major navigation or guidance components, there may not be sufficient CPU resources to execute every desired control rate without 'starving' the CPU of execution bandwidth needed for other tasks. There is no hard-and-fast rule for what percentage of the CPU the GN&C control algorithms are allowed; however, a good guideline is that they are allowed approximately 10 % of the available bandwidth. If not, either the control rate must be reduced (requirements will not be met) or the models can be executed in a multitasking environment where the control aspect in GN&C continues to execute at the control rate while the navigation, guidance, and other lower-priority functions within the GN&C execute at lower rates (typically 1 Hz and as much as 10 Hz for guidance, again depending on the mission requirements). In general, the model sub-rates typically have a modulo relationship that is evenly divisible by the control rate and not some arbitrary rate. This, of course, means that the model must be designed with this multi-rate requirement in mind. However, purely from a software perspective, it means that corresponding wrapper code must gather input data for the different rates, execute those sub-rates as separate tasks or threads, wait for them to complete execution, and gather the outputs. Again, the priority of the different tasks is set to guarantee completion of those tasks within the time period allocated. Generally, the high-rate task must be set to a high priority, but the lower-rate tasks can be set to much lower priorities. However, if the model is dominated by a single rate, and only a few blocks execute at a slower rate, multitasking can actually degrade performance. In such a model, the overhead incurred in task switching can be greater than the time required to execute the slower blocks. In this case, it is more efficient to execute all blocks at the dominant rate.

16.9.2.3 Asynchronous Model Execution

It is also possible to have asynchronous execution of some model component, particularly algorithmic components that may take several seconds or even minutes to complete. An example of this might be an orbit integrator that draws in large amounts of data that are subsequently integrated, resulting in a lengthy process simply because of the volume of data involved and the iterative nature of the algorithm. In this case, the independent task is assigned one of the lowest priorities within the CPU and is executed whenever the data it needs are available. Once complete, perhaps several minutes later, the results are forwarded to the GN&C model's function that requires the input at some modulo rate boundary, typically 1 Hz.

16.9.3 Managing Inputs

Typical GN&C sensors include IMUs, star trackers, and Sun sensors, to name but a few. Although they all execute at differing rates, the common aspect between them is that the inputs are read from the low-level device driver, such as RS-422, SpaceWire, or MIL-STD-1553, at the native rate for that device; the data are then time-tagged or otherwise associated with a time, given a ‘health-check’, and then provided as input within the same time boundary.

IMUs typically provide body rate and acceleration data at a cadence at least twice the control rate (typically at a rate of 100 or 400 Hz) and are not typically synchronized to the model execution rates. For example, for a notional 50-Hz control rate, two 100-Hz IMU samples are read, their data are then typically checked for staleness, and the IMU status flags are read to ensure that there is no internal fault; the data are then time-tagged if a time tag is not already provided by the device itself, supplied as input to the control task, and likely buffered for use by the lower rate navigation and guidance tasks as well as diagnostic telemetry.

An ST that provides inertial attitude knowledge as well as rate data is, however, typically synchronized to the control rate. For a 10-Hz tracker, attitude data, rate data, and device status are read once every fifth 50-Hz control cycle. The data are typically checked for staleness, internal status codes are checked for internal errors, and the data are time-tagged and buffered. Sometimes ST rate data are provided to control as a low-rate measure of body rates in case the IMU cuts out or simply as a secondary check against the IMU-detected rates, but generally, all 10 samples and associated health checks are provided to the guidance task as well as being buffered for diagnostic telemetry.

A low-rate sensor such as an SS is typically read at 1 Hz but in exactly the same way as the higher-rate sensors. That is, the data are read, health-checked, time-tagged, and provided, in this case, to guidance to typically verify spacecraft attitude or spin rates.

16.9.4 Managing Outputs

Typical actuators include RWs, torque rods, and attitude control thrusters as well as main propulsion engines.

Reaction wheel assemblies (RWA) are typically controlled at the control rate of the GN&C system. At the completion of the control task execution period, the RWA torque is read, sometimes formatted as required by the RWA and intermediate controlling hardware, immediately output within the control period, and finally buffered for the purposes of diagnostic telemetry.

Attitude control thrusters, which can be composed of thrusters of varying force capability, are also typically

controlled at the control rate. At the completion of the control task execution period, the on/off status of individual thrusters is read, often formatted as required by the intermediate controlling hardware, immediately output before the end of the control period, and finally buffered for the purposes of diagnostic telemetry. Before using the thrusters, there is usually a series of ‘thruster preparation’ commands to open various valves, turn on heaters, and so on, which are typically executed as part of a time-tagged command sequence that is outside of the GN&C subsystem and generally managed by the C&DH subsystem. This is particularly true of propulsion maneuvers, which are designed to impart a propulsive force along a specific direction, resulting in a change in spacecraft velocity (ΔV) along a particular trajectory. However, during the burn itself, control of the propulsion system (such as valves and heaters) may require greater control than that offered by an open-loop control via a sequence of time-tagged commands. In such a case, the GN&C typically is responsible for real-time valve switching as well as monitoring of the internal status of the propulsion system throughout the duration of the burn.

Torque rods or torque coils are devices that are typically used in low-Earth orbit in order to manage the attitude and momentum of the spacecraft by taking advantage of the Earth’s magnetic field. There are typically three that are configured orthogonally and controlled at a 1-Hz rate. Magnetic force is applied and controlled by switching the polarity of the individual rods based on a magnetic field model within the guidance or navigation components as well as inputs from a corresponding magnetometer sensor; this requires that the torque rods or coils be off during the reading of the ambient magnetic field that surrounds the spacecraft.

16.10 MESSENGER Case Study

To illustrate a full flight software architecture, the following section outlines the MERcury SURFACE, SPACE ENvironment, GEochemistry, and RANGING (MESSENGER) spacecraft to provide an example of the general principles described in preceding sections [7].

16.10.1 Spacecraft Overview

A simplified block diagram of the MESSENGER spacecraft is shown in Fig. 16.6. The spacecraft has two fully redundant integrated electronic modules (IEM), which contain the spacecraft bus processors. Each IEM contains an MP and a fault protection processor (FPP). These are RAD6000 processors, which execute the flight software applications. The flight software is implemented as C code that operates under the VxWorks RTOS.

16.10.2 Main Processors

The MP software implements all C&DH and GN&C functionality in a single flight-code application. Only one MP is designated ‘active’ or ‘primary’ and executes the full MP flight application. The ‘redundant’ or ‘backup’ MP typically remains unpowered because of MESSENGER mission power constraints, and serves as a cold spare. The backup MP, if powered, remains in boot mode and supports rudimentary command processing and telemetry generation for the purpose of reporting the health status of that processor, and to support uploads of code and parameters to EEPROM. It operates as an RT on the 1553 data bus. The primary MP serves as 1553 BC and manages all communication with devices on that bus.

C&DH functionality in the primary MP includes

- Uplink and downlink management using CCSDS protocols.
- Command processing and dispatch to other spacecraft processors and components.
- Support for stored commands (command macros) and time-tagged commands.
- Management of an 8 Gbit SSR using a file system.
- Science data collection.
- Image compression.
- Telemetry generation.
- Memory load and dump functions.
- Support for transmission of files from the SSR on the downlink using CFDP.

The uplink and downlink functions include control of two transponders via the 1553 bus. C&DH software also collects analog temperature data from temperature remote input output (TRIO) sensors, via a 1553 interface to the power distribution unit, and implements a peak-power tracking algorithm to optimize charging of the spacecraft battery via the power distribution unit interface. To support operational autonomy actions, the MP incorporates the same autonomy rule engine that is implemented in the FPP software. A number of C&DH functions interface to the spacecraft through an interface card that is in the IEM. For example, the uplink/downlink data buffers are on that card. The interface card also allows critical hardware commands to be sent from the ground or the FPP to force resets of spacecraft processors.

GN&C functionality in the primary MP includes attitude determination and attitude control, support for numerous spacecraft pointing modes, active control of solar panel orientation with respect to the Sun, momentum management, and two spacecraft safety modes: Safe Hold and Earth Acquisition. Safe Hold mode maintains a fixed power-positive pointing angle with respect to the Sun, with an antenna pointing to the Earth. Earth Acquisition mode addresses loss of attitude or time knowledge and is capable of pointing to the Sun using SS inputs, while

rotating about the Sun line to establish communications with Earth.

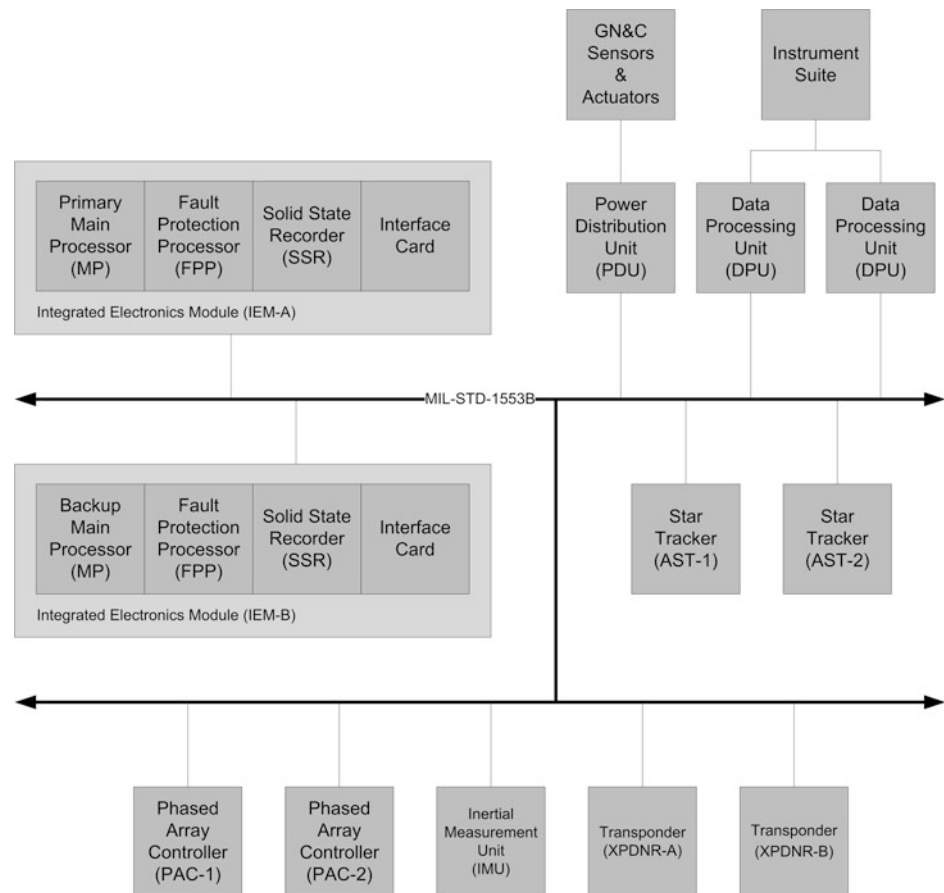
The GN&C software controls attitude with RWs or thrusters. Thrusters are used for trajectory correction maneuvers, including Mercury orbit insertion, and to reduce spacecraft momentum. GN&C sensors include two STs, an IMU that is internally redundant, and digital solar attitude detectors (DSADs, usually referred to as SSs). GN&C software also manages a precise ephemeris and a coarse ephemeris to support attitude determination. GN&C software interfaces to the STs, IMU, and phased-array controllers (to actively steer the high-gain phased-array antennas to point to Earth) via the 1553 bus. The remaining GN&C sensors and actuators have custom hardware interfaces to the PDU, which, in turn, provides a command and telemetry interface to those devices via the 1553 data bus. These devices include the RWA, propulsion system, SSs, and solar-array drive electronics (SADE).

The primary MP interfaces to two data processing units (DPU) and the two FPPs via the 1553 data bus. The DPUs provide the interface to all other instrument processors. GN&C software in the MP passes data to the primary DPU to route attitude data to the imager and laser altimeter instruments, and to actively steer the imager pivot motor.

16.10.3 Fault Protection Processors

The two FPPs are on unswitched power so that both are always powered, although a critical hardware command allows ground controllers to hold either in boot mode if needed. Each FPP executes an identical flight code application that supports a command and telemetry interface to the MPs via the 1553 data bus. The main purpose of each FPP is to perform fault detection and to isolate fault correction responses within these processors. Each FPP implements an autonomy rule engine, which accepts uploadable health and safety rules that can operate on data collected from the 1553 data bus or a state message transmitted by the primary MP. In addition to being an RT on the 1553 bus, each FPP serves as a 1553 bus monitor to collect spacecraft data that can be monitored by autonomy rules. The rules are expressed in reverse Polish notation, and the action of each rule can dispatch a command (or a series of commands from a stored FPP macro) to the primary MP for subsequent execution by the MP to correct faults. Fault correction can include actions such as switching to redundant components, demotion to Safe Hold or Earth Acquisition mode, or shedding power loads. Additionally, the FPPs have a custom serial interface via the interface card to the PDUs in order to receive PDU critical status updates or send special commands in the event of loss of 1553 bus communications or a failed MP. The PDU command interface allows the FPPs to swap the BC

Fig. 16.6 MESSENGER spacecraft block diagram



functionality between MPs, reset the MP in its own IEM, select which of two stored flight applications either MP loads and executes, or power on and switch to the redundant MP and declare it primary.

16.10.4 MP and FPP Boot Software

The primary function of the boot software in the MPs and FPPs is to perform processor and memory diagnostic tests, processor and hardware backplane initialization functions, and verification-load-launch of one of two stored flight-code applications (as designated by hardware signals that can be set by ground command via the critical command decoder). The primary MP and both FPPs always progress from boot code to execution of the designated flight software application in RAM. Only the backup MP, if powered on, remains in boot mode. The boot software for the backup MP includes support to maintain communications with the primary MP by acting as an RT on the 1553 data bus.

16.10.5 Spacecraft Instruments

The MESSENGER spacecraft includes two DPUs. Only one DPU is designated 'primary' and interfaces to the suite of

other instrument event-processing units (EPUs) via serial interfaces. It communicates with the MP via the 1553 data bus and is the processor that provides all science data and telemetry to the MP and accepts all instrument commands from the MP. The backup DPU is typically unpowered because of MESSENGER power constraints. Only the primary DPU interfaces to the seven instrument processors (EPUs). The backup DPU can take over as primary to control the EPUs, but there is no ability to control some from one DPU and the remainder from the second DPU. The MESSENGER instruments include

- Mercury dual imaging system (MDIS). The MDIS instrument hardware is controlled by software that is part of the DPU flight application.
- Magnetometer (MAG).
- X-ray spectrometer (XRS).
- Energetic particle and plasma spectrometer (EPPS).
- Gamma-ray and neutron spectrometer (GRNS).
- Mercury laser altimeter (MLA).
- Mercury atmospheric and surface composition spectrometer (MASCS).

The DPUs and the EPUs for all instruments except MLA each use an RTX2010 processor, and the flight software is implemented in the Forth language, which is native to that processor. The processor for the MLA instrument uses an

Intel 80C196KD processor, with the CMX operating system and software written in C.

The flight software for the RTX2010-based DPU and EPUs shares a core of ‘common’ Forth code that implements standard functionality such as communications interfaces, memory management, and command macro storage and processing. In addition, each of the DPU and EPU flight applications has instrument-specific Forth code to implement the functionality unique to each instrument.

Time is distributed to the instruments via a hardware interface that provides the 1 pulse-per-second (PPS) clock to the DPU, which passes it on to each EPU. The primary MP uses a 1553 bus message to distribute MET to the DPU, which subsequently passes it to the EPUs so that they can synchronize MET with the 1 PPS. The 1 PPS clock comes from the IEM interface card, which offers a choice of a precision oscillator (needed to meet 1-ms time correlation requirements for MDIS) or a coarse oscillator. Only one instrument, MDIS, has an independent hardware interface directly to the IEM. Images collected by the MDIS go directly to a buffer on the interface card via a high-speed serial interface, so that MP flight software can manage the storage of those images on the SSR file system.

16.10.6 Summary

Spacecraft flight software consists of multiple elements on-board a spacecraft. On the main computer, typical applications include the boot software which initializes the hardware; C&DH which handles commanding, telemetry, and other data processing tasks; and GN&C which controls the spacecraft’s attitude, thrusting, and other guidance tasks. Other flight software includes applications on science payloads, power systems, and transceiver and transponder modules. Due to its critical nature, software engineers follow a strict development process when developing flight software. An example development process is the Waterfall model which has an incremental development flow from concept to requirements to design to implementation to testing to delivery. Once in-flight, the software team works with mission operations to ensure full functionality of the software through support and uploads of new software parameters and code images.

16.11 Further Reading

For further information on the current state of spacecraft flight software research and concepts, consult the proceedings of the annual Workshops on Spacecraft Flight Software located on the website <http://www.flightsoftware.org/>.

Books and publications specifically relating to spacecraft flight software are rare. However, general embedded software techniques and practices are broadly applicable to flight software development. Books in this area include the following:

Barr, M., Massa, A., *Programming Embedded Systems with C and GNU Development Tools*, O’Reilly Media, Sebastopol, CA, 2007.

Catsoulis, J., *Designing Embedded Hardware*, O’Reilly Media, Sebastopol, CA, 2005.

Koopman, P., *Better Embedded System Software*, Drumndrochit Press, 2010.

Noergaard, T., *Embedded Systems Architecture*, Newnes, Burlington, MA, 2005.

Simon, D. E., *An Embedded Software Primer*, Addison-Wesley, Boston, 1999.

White, E., *Making Embedded Systems*, O’Reilly Media, Sebastopol, CA, 2012.

For further information on the history of spacecraft computer and computing:

O’Brien, F., *The Apollo Guidance Computer: Architecture and Operation*, Springer/Praxis, Berlin, 2010.

Tomayko, J. E., *Computers in Space: Journeys with NASA*, Alpha Books, Indianapolis, Indiana, 1994.

For more information on the MESSENGER mission:

The MESSENGER Mission to Mercury, Edited by D. L. Domingue and C. T. Russell, Springer, New York, 2007.

References

- O’Brien, F., *The Apollo Guidance Computer: Architecture and Operation*, Springer/Praxis, Berlin, 2010.
- Tomayko, J. E., *Computers in Space: Journeys with NASA*, Alpha Books, Indianapolis, Indiana, 1994.
- “IEEE Standard Glossary of Software Engineering Terminology, 610.12-1990,” Institute of Electrical and Electronics Engineers, 1990.
- “Space Packet Protocol,” CCSDS 133.0-B-1, Consultative Committee for Space Data Systems, September 2003.
- “Packet Telemetry,” CCSDS 102.0-B-5, Consultative Committee for Space Data Systems, November 2000.
- “CCSDS File Delivery Protocol,” CCSDS 727.0-B-4, Consultative Committee for Space Data Systems, January 2007.
- Leary, J. C., Conde, R. F., Dakermanji, G., Engelbrecht, C. S., Ercol, C. J., Fielhauer, K. B., Grant, D. G., Hartka, T. J., Hill, T. A., Jaskulek, S. E., Mirantes, M. A., Mosher, L. E., Paul, M. V., Persons, D. F., Rodberg, E. H., Srinivasan, D. K., Vaughan, R. M., Wiley, S. R., “The MESSENGER Spacecraft,” *The MESSENGER Mission to Mercury*, Edited by D. L. Domingue and C. T. Russell, Springer, New York, 2007.

Extending human activities to outer space has been a major target of space engineering from its inception. We have long dreamed of space flight, been curious about the origin of the universe, our solar system, and life on Earth [1]. Even with recent discoveries of many extra-solar planets, Earth remains a uniquely habitable planet. Living organisms themselves have modified the terrestrial environment by their activities, and helped to maintain its habitability. In a spaceship or at outposts built on extraterrestrial bodies, life support engineering must create an environment approximating the Earth's biosphere. As crew sizes and system operation times are increased, the recycling of materials, or 'closing the loop', will gain an economic advantage over open loop systems. Although the most critical index for life support engineering is ensuring the survivability of the space crew, life management (i.e., quality of life) is also essential to the fulfillment of human needs.

17.1 Environmental Control for Life Support

Our terrestrial biosphere has evolved over the long history of the Earth, and living organisms are well adapted to this environment. Each species can sustain itself within a certain range of environmental parameters, such as atmospheric composition and thermal conditions. Space life support systems should be equipped with the capability to synthesize and maintain the environment in this range for the space crew and companion living organisms.

M. Yamashita (✉)
Institute of Space and Astronautical Science (ISAS), Japan
Aerospace Exploration Agency (JAXA) (Emeritus), Tokyo, Japan
R. M. Wheeler
John F. Kennedy Space Center, National Aeronautics and Space
Administration (NASA), Florida, USA

17.1.1 Cabin Air

A space cabin is principally a pressurized structure, maintaining an atmosphere inside. Gas that leaks from the cabin must be supplemented from a reservoir. The capacity of the reservoir and the resupply frequency is determined by the rate of consumption by crew and leakage from the cabin. Leak rates can be reduced by lowering the cabin pressure. Determining the optimum cabin pressure is a trade-off between human physiology in a hypobaric environment and the engineering to reduce the leak rate. The habitable upper-altitude for humans with the normal composition of air is 4,000 m above sea level for ordinary life, although it is possible to operate at 8,000 m for a short duration. Tolerable ranges of total pressure and partial pressure of oxygen are shown in Fig. 17.1 [2]. The oxygen percentage must be higher in hypobaric conditions in order to prevent hypoxia, and since hyperoxic conditions are also known to induce a health risk, oxygen levels should be kept lower than this limit. Furthermore, it is important to keep the oxygen partial pressure and percentage below certain threshold to minimize fire risks. In order to prevent ignition and propagation of combustion, oxygen should be diluted by inert gas species.

Decompression syndrome during extra-vehicular activity (EVA) is another factor that influences the choice of cabin air pressure. As EVA suits are as low as 30 kPa with pure oxygen, a period of pre-breathing at an intermediate pressure should be performed prior to an EVA. However, if the decompression rate is too fast, bubbles of inert gas (mainly nitrogen) are formed intravenously. To avoid the formation of these bubbles, a well-prepared pre-EVA protocol should be followed, with a staged decrease in breathing pressure. Pre-breathing pure oxygen is known to be effective in avoiding bubble formation. One reason that the pressure of the EVA suit differs from cabin pressure is to maintain flexibility of the mobility joints of the EVA suit under vacuum. A possible engineering approach for achieved this

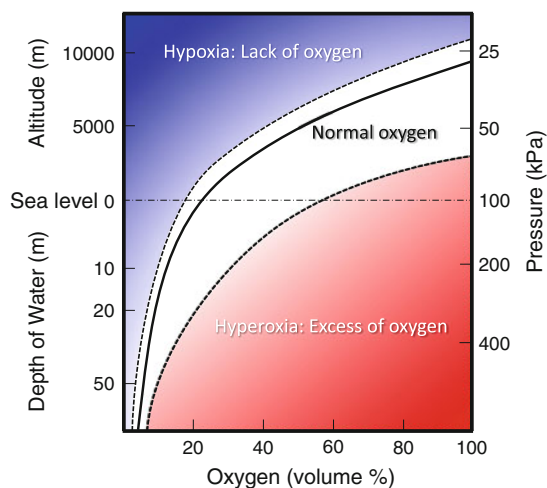


Fig. 17.1 Hypoxia and hyperoxia danger zones at different partial pressures of oxygen and at different total pressures

flexibility could be to provide power assisted functions to the EVA suit (Fig. 17.2).

Carbon dioxide and water vapor are minor components of cabin air. A high carbon dioxide partial pressure causes headaches and nausea. The allowable carbon dioxide partial pressure depends on the duration of exposure: 2 kPa maximum for 1 h and 0.7 kPa for 7–180 days, respectively [3]. Humidity is maintained at an appropriate level mainly for crew comfort. Common toxic gas species that can be released into the space cabin are listed in the Spacecraft Maximum Allowable Concentrations for Airborne Contaminants (SMAC) [4]; the limits of common compounds are given in Table 17.1 [5]. These hazardous gases can be released by materials or created during accidental fires and other events. Use of such high-risk materials should be avoided where possible. The gases listed in Table 17.1 should be monitored, and automatic alarms activated when they exceed prescribed limit.

As the natural convection of gas as a result of buoyancy is suppressed by the absence or reduced level of gravity, cabin air should be circulated to prevent uneven distribution of the gas species that are consumed or produced inside the space cabin. The species of concern are oxygen, carbon dioxide, and minor metabolites such as ketones, organic acids and esters. Forced circulation must sweep all the volume in the cabin without leaving stagnant areas. Gradients of concentration of gas species can form along the path of the ventilation line. Ideally, cabin air should be sampled from each compartment with due consideration of possible gradients of concentration. The monitoring of major components (oxygen, carbon dioxide and humidity) is essential for house-keeping of the life support functions. The SMAC gas species can be analyzed by gas chromatograph/mass spectrometer (GC/MS), Fourier transform infrared spectrometer (FTIR) and other analytical devices, which are tested and operated

on-board. Off-line analysis of trace species can also be conducted on the ground for detailed evaluation.

17.1.2 Water Management

Water is a critical resource for drinking, rehydration, hygiene, and medical use. On the Space Shuttle, the H_2O_2 fuel cell system for electric power generation produced water as its byproduct. Electric power on-board the International Space Station (ISS) is generated by solar cells. Since water is the most heavily consumed item in life support, it is recycled on the ISS at the rate of 3 kg per crewmember per day for physiological needs, plus potentially another 26 kg per crewmember per day for hygiene, flushing, laundry, and dishes. The palatability of this processed water is a critical factor for crew psychology, and must be considered in the life support engineering.

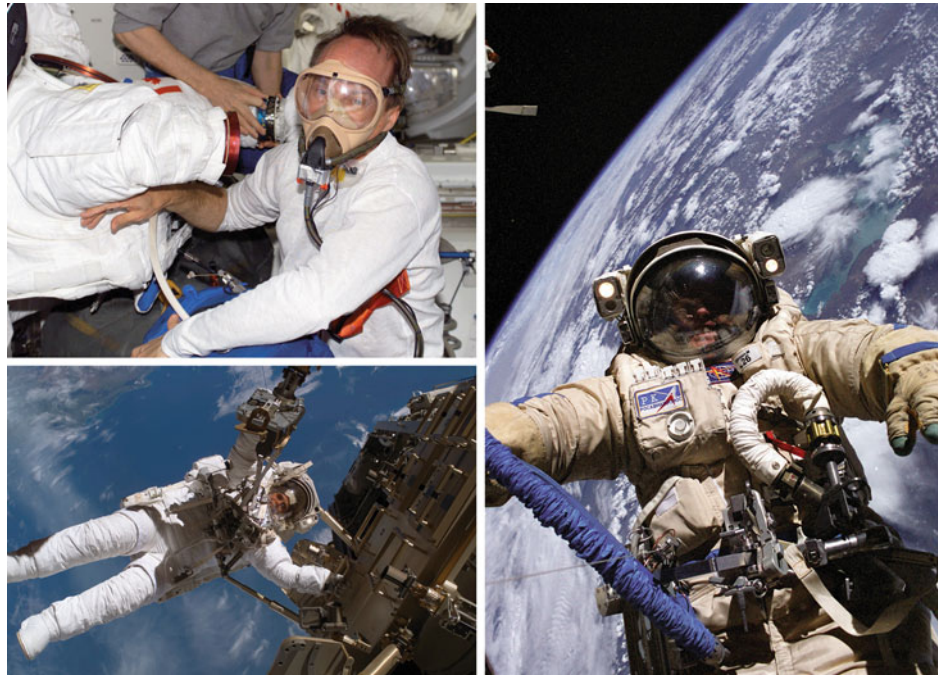
Water quality is monitored on-board to check whether it meets the requirements for potable water. The allowable limits of inorganic elements and organics are summarized in Table 17.2. Current specifications by agencies such as NASA require microbial contamination to be less than 50 colony formation units (CFU) per millimeter for bacteria, and not detectable per 100 ml for coliform bacteria, fungus, and parasitic protozoa. Real-time monitoring should be deployed on-board the spaceship for the defined pollutants and contaminants. If the analysis toolset is insufficient, then total organic carbon (TOC) represents the overall contamination level of water.

Microbial monitoring is routinely conducted by off-line analysis of collected samples retrieved on the ground. An autonomous on-board system to identify microbial species and analyze their population size is being developed for monitoring cabin cleanliness, whereby suspended bacteria are stained in a micro fluid device and detected by a fluorescence microscope. Biofilms often form inside water storage systems, and can be a serious contamination source. Iodine or silver ions are added to such systems as a biocide to suppress the growth of microbial populations. The concentrations of these chemicals should be continuously monitored and kept within a defined range. Alternative anti-bacterial measures should be prepared for the possible mutation of resistant bacteria after long-term exposure to space radiation.

17.1.3 Illumination

Illumination affects the visual perception of the crew when conducting tasks in the cabin and laboratory. The amount of light and its color spectrum are important factors for the psychology and physiological performance of space crew. If

Fig. 17.2 EVA suit, a minimum-scale life support system. Astronaut Steven G. MacLean, of the Canadian Space Agency, shown in the midst of a pre-breathing exercise in the quest airlock of the International Space Station in preparation for an EVA (*top left*). ESA astronaut Christer Fuglesang, during an EVA [*bottom left* NASA image ISS014-E-09795 (14 Dec. 2006)] and NASA astronaut Edward M. (Mike) Fincke, wearing a Russian Orlan spacesuit, at the International Space Station (*right*). Image NASA



a person is kept in complete darkness for a prolonged period, his/her mental status is deeply depressed and harmed. White light of an appropriate color temperature can be provided by electric lamp, such as incandescent, fluorescent or light emitting diodes (LED). The luminance, power efficiency, and lifetime of lamps are engineering factors to be considered when designing lighting systems. The required minimum lighting level depends on the types of task to be performed. For night lighting during the sleep period, the light intensity is lowered.

Natural sunlight can be introduced into the space cabin through windows, or through a solar light collecting system. Sunbathing is one application of such a system. Strong light is an effective cue for maintaining normal circadian bi-rhythms. When natural sunlight is used for this purpose, the shorter ultraviolet (UV) portion should be filtered out, as it is by the Earth's atmosphere, to avoid harmful effects on the crew. If the heat load must be suppressed, it is better to remove the infrared (IR) part of the light. Solar light collection system focuses incident sunlight into its light guide by either a lens or mirror. Color aberration of the lens and/or transmittance qualities of the materials can remove harmful UV and IR rays, while allowing the visible light portion into the space cabin, if the lens optics are adequately adjusted.

17.1.4 Sound

Sound is a useful medium to perceive and gain information from the surroundings, and to communicate with other

crewmembers or ground personnel. The acoustic environment in a space cabin during the mission phase is moderate compared to the noisy launch and descent periods. Even though the sound pressure is less during the mission phase, annoying noises should be suppressed in the crew's living cabin both to avoid distraction during work tasks and to allow for better sleep and relaxation. The criteria for the sound environments are defined for the hearing frequency range. However, both infrasonic and ultrasound noise can induce indeterminate symptoms even when they are hard to hear. Ventilation fans and other mechanical and electrical devices are possible sources of infrasonic noise and their acceptability should be carefully evaluated for crew health.

17.1.5 Space Radiation

The space environment is discussed in detail in [Chap. 3](#), and from a human effects perspective here.

17.1.5.1 Characteristics of Space Radiation

Radiation is categorized as either non-ionizing or ionizing. Electromagnetic waves from radio frequency to visible light are non-ionizing radiation, since their energy is insufficient to excite molecules to an ionized state. If the effects of radiation are limited to the deposit of thermal energy, terrestrial organisms can withstand a heat flux of 1 kW/m^2 , equivalent to the energy density of terrestrial incident solar light. However, physiological responses other than thermal effects are possible for longer electromagnetic waves.

Table 17.1 Limits of common toxic gas species

Compound	Limit (mg/m ³)
Methanol	0.5
Ethanol	5
2-Propanol	5
2-Methyl-2-propanol	5
N-butanol	5
Ethanal (acetaldehyde)	0.5
Benzene	0.1
Xylenes	10
Methyl benzene (toluene)	3
Dichloromethane	0.5
Dichlorodifluoromethane (Freon 12)	10
Chlorodifluoromethane (Freon 22)	5
Trichlorofluoromethane (Freon 11)	10
1,1,1-Trichloroethane	1
1,1,2-Trichloro-1,2,2-trifluoroethane (Freon 113)	5
N-hexane	5
N-pentane	10
Methane	180
2-Methyl-1,3-butadiene	10
Propanone (acetone)	1
2-Butane	3
Hydrogen	10
Carbon monoxide	2
Hexamethylcyclotrisiloxane	10
Trimethylsilanol	3
2-Butoxyethanol	1
Trifluorobromomethane (Halon 1301)	10
Carbonyl sulfide	0.5
Acetic acid	0.5
4-Hydroxy-4-methyl-2-pentanone	1

The depth of energy penetration through the body surface depends on wavelength. Consequently, the exposure limit for electromagnetic waves is typically defined with reference to its frequency (Hz).

Shorter electromagnetic waves, such as UV, X-rays and gamma-rays, are capable of ionizing molecules by excitation above the ionization threshold. Incident particles with energy sufficient to ionize molecules are also termed 'radiation'. Electrons (beta-rays), neutrons, protons and other ionized atomic nuclei, including helium nuclei (alpha-rays), are major components of this radiation. High-energy phenomena in the Sun and the Milky Way galaxy, which are not yet fully understood, emit energetic photons and can accelerate charged particles. These are termed cosmic rays. On Earth's surface, the thick atmosphere and the strong geomagnetic

Table 17.2 Potable water physiochemical limits [2]

Chemical	Limit (mg/L)
Ammonia	1
Antimony	2
Barium	10
Cadmium	0.022
Manganese	0.3
Nickel	0.3
Silver	0.4
Total iodine	0.2
Zinc	2
Total organic carbon	3
Acetone	15
Alkylamines (di)	0.3
Alkylamines (mono)	2
Alkylamines (tri)	0.4
Benzene	0.07
Caprolactam	100
Chloroform	6.5
Di (2-ethylhexyl) phthalate	20
Di-n-butyl phthalate	40
Dichloromethane	15
Ethylene glycol	4
Formaldehyde	12
Formate	2,500
2-Mercaptobenzothiazole	30
Methanol	40
Methyl ethyl ketone (MEK)	54
Phenol	4
n-Phenyl-beta-naphthylamine	260

field that deflects the trajectory of charged particles, reduce the incidence of cosmic rays. Since those shielding mechanisms are not present in outer space, astronauts are directly exposed to the harsh space radiation. Space radiation is characterized by a high flux of energetic photons and heavy nuclei, which are rare on the Earth's surface. Nuclei with high atomic numbers (*Z*) and energy, mostly galactic in origin, are abbreviated as HZE.

High energetic electrons and protons, major components of the solar wind, are trapped in the van Allen belts by the Earth's magnetic field (see [Chap. 3](#)). As discussed in [Chap. 4](#), an altitude of 300–500 km is high enough to avoid the very worst drag caused by the residual atmosphere whilst also providing some beneficial shielding from the residual atmosphere and remaining below the radiation belts. As such, for crewed activities an altitude of 300–500 km is preferred in order to limit the exposure of the crew to radiation.

17.1.5.2 Biological Effects of Space Radiation

The biological effects of radiation differ depending on the type of radiation and its energy level. Aside from thermal effects, radiation can cause chemical bonds to break and form free radicals, peroxides, and other species, which can affect biochemical processes in cells. The relative biological effectiveness (RBE) of cosmic rays is less known, but is an important factor in defining the exposure dose criteria for crewed activities in space. A higher occurrence of double-strand breaks in chromosomes is specifically caused by HZE. Radiation damage to cytoplasm and cellular organelles has also been studied in microscopically controlled irradiation experiments.

The dose rate dependence is considered for the biological effects of radiation. Acute effects on human health are induced at doses higher than 0.5 Sv, where the Sievert (Sv) is a unit of dose to evaluate biological effects, 1 gray (Gy) multiplied by RBE; the gray is a SI unit of dose measured by energy deposit per unit mass. Radiation injury is caused stochastically even at a very low dosage. Taking ambiguity of scientific assessment into account, the radiation exposure limit for the public is set by the International Commission on Radiological Protection (ICRP) to 1 mSv/year from artificial sources. However, it is noted that aircraft crew (and frequent passengers) who will spend many hours per year in the upper troposphere, around 10 km altitude, can get an extra dose of >2 mSv/per year. The most serious risks to humans are oncogenesis (or carcinogenesis) and genetic effects. Cell damage from radiation is propagated to neighboring cells by the diffusion of signal transmitters. This side effect, which amplifies the number of damaged cells beyond that of the cells that are directly hit, results in a non-linear dose dependence at the low dosage end. On the other hand, there is a chance to repair damaged DNA (deoxyribonucleic acid) during the cell cycle. This permits a higher integrated dosage, if the dose rate is low enough to permit the functioning of the repair processes.

Radiation dose rates in the ISS can average 1 mSv/day during the calm period of solar activity. The natural radiation dose rate on the Earth's surface is typically 2.4 mSv/year, however sizeable population groups receive up to 10–20 mSv/year. Biological systems are evolutionarily adapted to this natural dosage. Staying in the ISS far exceeds the international criteria for exposure: 50 mSv for one year, or 100 mSv over 5 years. For astronauts, lifetime exposure limits are 1,200 mSv for male astronauts, whose first space flight occurs after the age of 40 and 600 mSv for young female astronauts. There are no criteria for pregnant astronauts, since such cases would not be permitted. Intensive crew health care and life-long monitoring might mitigate the health risk to astronauts. Epidemiological studies on commercial airline crews have shown they are exposed to a ten times higher dose of radiation during flights at an altitude of

10 km, compared to people remaining on the ground. However, although skin and breast cancers are higher among airline crews, the total number of cancer cases is less than in the control groups. This may be attributed to differences in available health care.

The technology for monitoring radiation in space has been developed intensively in parallel with the development of human space flight. In addition to passive dosimetry such as plastic plate and thermoluminescent dosimeter (TLD), real time dosimetry is conducted in orbit. Phantom measurement of a synthetic human body embedded with dosimeters is a key to assessing absorption coefficients of biological tissue for space-specific radiation quality and for mapping the radiation dosage in each part of the body.

17.1.5.3 Protective Measures Against Space Radiation

The radiation environment around Earth is largely dominated by solar activity. Charged particles, mainly protons, are accelerated by the magnetic reconnection formed over a solar flare. Ejections of large numbers of accelerated particles are called solar particle events (SPE). A high dosage, up to 1 Gy, is typical in a short period of a few days. By observing activity on the Sun's surface, this ejection of dense energetic particles can be predicted. Upon receiving a high-severity SPE forecast alert, an EVA should be terminated and the possibility of an emergency return to Earth should be considered. In moderate SPE events, crewmembers could hide in the most heavily shielded part of the space cabin.

Shielding is a general measure against space radiation. The Earth's atmosphere is equivalent to a 10 m thick layer of water shielding against radiation. The interaction of high-energy particles with shielding materials generates a shower of secondary radiation, including fast neutrons. Such secondary radiation is one characteristic feature of space radiation experienced inside the space cabin. An insufficient layer of passive shielding materials results in numerous small damages spread over a wide area ('shot gun'), instead of one large localized area of damage ('cannon'); see Fig. 17.3.

Similar to the shielding produced by the geomagnetic field, active shielding is a concept of forming a strong magnetic field or a steep electric field around a spaceship in order to deflect or decelerate incoming cosmic ray particles. In addition to assessing the technical feasibility of building such a shielding field, any negative biological effects caused by this shielding field should be carefully evaluated. There are as yet few studies available on the health risks of a strong electromagnetic field, but neural activities and early development are known to be affected by exposure.

Development of anti-radiation medication might be a promising approach. Extremophilic organisms are tolerant to various stresses including radiation. They possess molecular

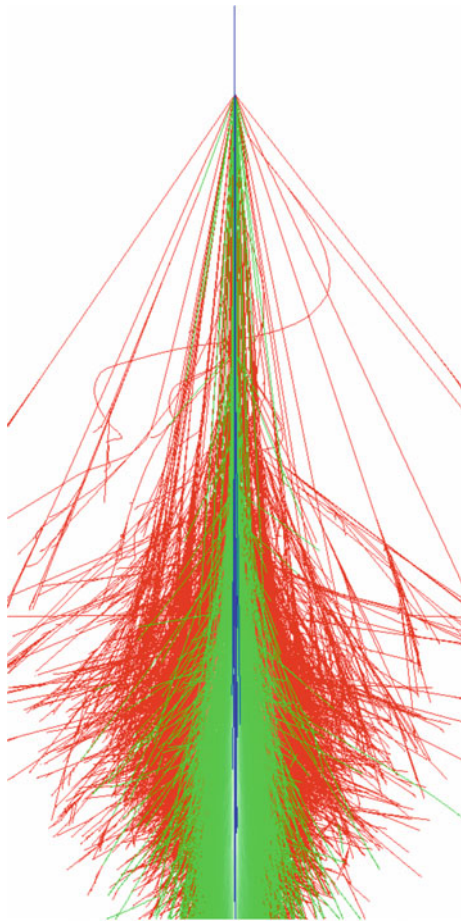


Fig. 17.3 Shower of secondary particles made by a high-energy particle interaction

machinery to repair damaged chromosomes. Over-expression of the repair-associated genes might effectively prevent and cure radiation damage. Suppression of oncogenesis (carcinogenesis) or enhancement of apoptosis for damaged cells may be another effective measure. Understanding the systems biology of living cells regarding these phenomena is therefore critical.

17.1.6 Biological Environment

Other living organisms are part of the space crew's environment. Even when strong controls are applied to prevent biological contamination of the cabin, the crew themselves cannot be sterilized and can therefore be a source of various microorganisms. Fungi and bacteria are found in the cabin, either airborne, in potable water, or on surfaces. Since sedimentation does not occur under microgravity, airborne microorganisms become dominant in the space cabin. Pathogenic bacteria, protozoa, and allergenic fungi spores should be monitored and controlled within the allowable

limit. Fungi that proliferate on cabin surfaces and inside air ducts may cause symptoms similar to Sick Building Syndrome. The space cabin system needs to be capable of disinfection and sanitization.

Protection of our planet involves evaluating the risk posed by potential but presently unknown extraterrestrial living creatures and organic compounds. Although such a risk by its very nature can never be specific, it requires assessing the chance of extraterrestrial organisms being transported to Earth and threatening terrestrial organisms and our ecology. However, extraterrestrial life and organic substances are transported to Earth naturally as well; cosmic dust containing organic substance(s) is estimated to enter the Earth's atmosphere at $4 \pm 2 \times 10^7$ kg/year. The inclusion of extraterrestrial life in cosmic dust or larger bodies cannot be ruled out. Certainly, space activities may increase the risk by shortening the travel time of extraterrestrial life, if any, and by protecting it from the harsh conditions otherwise experienced during the ejection and entry phases. Fine regolith particles brought into the space cabin after an EVA on an extraterrestrial body is recognized as a health risk, because of the highly active chemistry of such material.

17.1.7 Fire Safety

Fire is a crucial safety issue that has the potential to end in catastrophe. In order to prevent fire in the space cabin, the building materials and contents should be manufactured from fire-retardant substances. Their ignition temperature should be high, and any combustion should never propagate. The oxygen concentration of cabin air is maintained below 30 % to reduce the risk of fire propagation. Sources of hot spots and sparks should be carefully removed by design, and verified by testing. The toxic chemicals commonly produced by combustion can include carbon monoxide, hydrogen cyanide, and hydrogen chloride. In the fire detection system, forced airflow is needed to conduct (remove) these toxic gas species and smoke particles, because natural convection does not assist this detection under microgravity.

Should fire occur, electric power to that portion of the cabin should be removed. Fire is extinguished by the use of non-toxic chemicals that are appropriate for a confined environment. Any expended chemicals must be removed after a fire has been extinguished. To meet these requirements, a combination of carbon dioxide and water-based fire extinguishers have been selected for use in space. During a fire, the affected compartment is isolated and evacuated, if the criticality of the fire is high and a vacuum can be created to extinguish it. All compartments of the space ship must have established escape routes to a safe location in the event of an emergency evacuation. Cabin design and the layout of components must meet this safety requirement.

17.2 Materials Recycling

The space crew requires a continuous supply of life support consumables, e.g., oxygen, water, food, and other items. At the same time, the gaseous, liquid, and solid wastes that accumulate in the confined space cabin should be managed and processed adequately. For brief crewed space missions close to Earth, life support consumables have been brought up to space, and wastes have been returned or dumped to Earth. For longer missions in low Earth orbit (LEO), such as Mir and the International Space Station, physicochemical regenerative technologies have been used to recycle water from urine, since the penalty of increased mass is highest for water in the open loop system. For future space missions beyond LEO, such as interplanetary transit or outposts on extraterrestrial bodies, life support systems technology will have to be more advanced. For longer mission durations and greater distances from Earth, regenerative or materially closed loop life support technologies will be essential. Table 17.3 summarizes typical input rates of materials for human life support, and output of waste products for one person.

Regenerative life support greatly reduces the costs of resupplying consumables. However, the initial investment to build the closed regenerative system and the costs for operating it for a defined mission period should be examined in terms of the economy of life support. For longer, larger, and more widely ranging missions the estimated cost of regenerative systems becomes cheaper than the total sum of consumables for the open loop life support systems. Resources required to operate and control the life support system are additional factors in the trade-off between open and closed concepts. There might be an optimum degree of closure, or a good combination of the two. The ‘economics’ of life support systems and their operation can be measured by the mass, power, volume, and crew time expended. The use of self-sustaining, regenerative technology would not only reduce the costs of long-duration missions, but also reduce mission risks by increasing the level of autonomy. However, the top priority and criteria for the engineering of life support systems are robustness and survivability of the crew.

17.2.1 Physicochemical Recycling Systems

Waste regeneration by physicochemical processes is a well-established technology. From a systems engineering point of view, the system characteristics for physicochemical components are better defined than those of biological elements. Recycling of water and oxygen has been accomplished by physicochemical processes. A physicochemical regenerative system is shown in Fig. 17.4 [6].

17.2.1.1 Water Recycling

On the ISS, much of the clean water is supplied from Earth. Water collected at the heat exchange condensers is also sent through the treatment steps including filtration. Urine is either chemically stabilized for storage and then dumped or returned to ground, or it is sent to a distillation system. The distillate water is then sent through filtration beds. The water produced is treated with a biocide (silver) for potable water to reduce the risks of microbial contamination. Handling liquids like water under microgravity poses engineering challenges, especially for separating gases from water. For surface missions to the Moon or Mars, more conventional water handling might be applicable under one-sixth and one-third gravity, respectively.

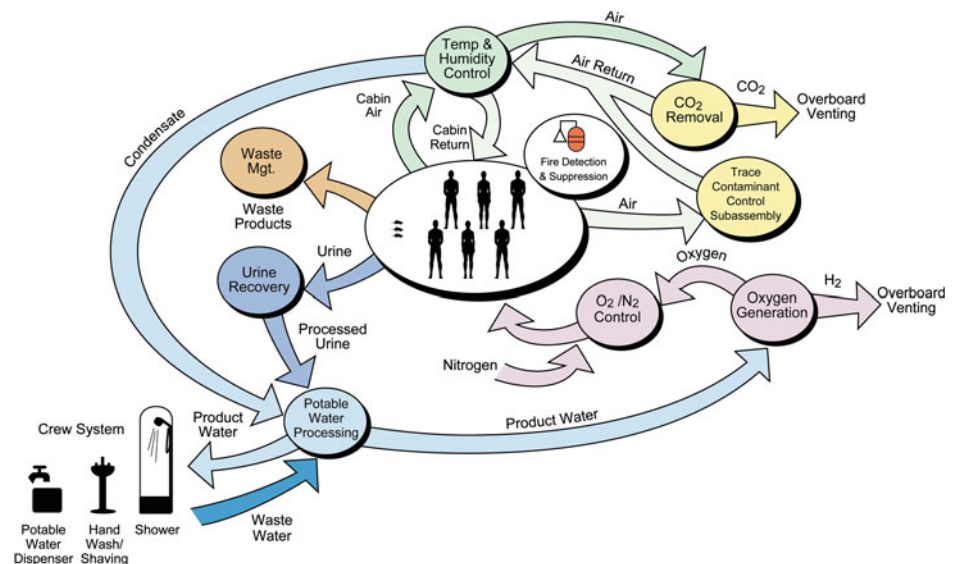
The mass and power required for distillation of wastewater ultimately needs to be reduced for water management in space. In principle, various approaches for water recycling could be used, such as reverse osmosis filtration technology. Wastewater containing urine is currently pretreated with strong acids to prevent volatilization of ammonia and to reduce microbial activity. These acids pose a safety risk to the crew, and alternative approaches should be considered for stabilizing the urine. The use of biological pretreatment to oxidize organics to CO_2 and convert nitrogenous compounds to nitrate or N_2 might be worth exploring. Further work on the efficacy of alternative biocides for converting and maintaining potable water is a high priority if biological treatment is included in the materials recycling system.

17.2.1.2 Gas Regeneration

Conventional approaches to CO_2 control in space cabin air are removal using strong alkaline reactants such as lithium hydroxide, and adsorption. Lithium hydroxide filters are limited to one-time use, and must be replaced when consumed. Adsorption to molecular sieve (zeolite) or other adsorbents can be used repetitively after appropriate revitalization of the adsorbents by raising the temperature and/or decreasing the pressure. Since both strong alkali and regenerative adsorbents for CO_2 will absorb water vapor as well, preprocessing is required to remove water from the cabin air before sending it to the CO_2 adsorbent beds. This also allows retrieval of water vapor for recycling. Water vapor on the ISS is also collected by heat exchangers that operate below the dew point of the cabin air. CO_2 retrieved from adsorption beds is reduced by hydrogen in Sabatier processors, producing methane and water as reaction products. The methane is vented, while the water vapor can be recycled. These CO_2 reduction technologies are a step toward increasing closure of the air regenerative loop, and they provide an additional source of water. Another method is the Bosch process, a catalytic reaction of CO_2 with hydrogen which produces carbon and water as its products.

Table 17.3 Life support requirements for one person; taken from the space station ECLSS Architectural Control Document, NASA SPP 30262

Inputs	Daily requirements		Outputs	
	(kg)	% of total mass	(kg)	% of total mass
Oxygen	0.83	2.7	Carbon dioxide	1.00
Food	0.62	2.0	Metabolic solids	0.11
Water (drink and food preparation)	3.56	11.4	Water	29.95
Water (hygiene, flush, laundry, dishes)	26.0	83.9	Metabolic/urine	12.3
<i>Total</i>	<i>31.0</i>		Hygiene/flush	24.7
			Laundry/dishes	55.7
			Latent	3.6
			<i>Total</i>	<i>31.0</i>

Fig. 17.4 Physicochemical regenerative environmental control and life support system diagram

Oxygen (O_2) is supplied from ground in the form of either compressed gas, or solid chemical oxygen generators, called oxygen candles. The solid O_2 generators are typically kept for emergency situations, and consist of chlorate and perchlorate compounds. Electrolysis of water has been used to generate O_2 on both Mir and the ISS. The hydrogen produced by water electrolysis can then be supplied to the Sabatier processor for reducing CO_2 , thereby further closing the air and water recycling loop.

17.2.1.3 Trace Contaminants and Dust Control

Since space cabins and surface habitats are tightly closed environments, high concentrations of airborne trace contaminants or dust are accumulated. Because these can cause safety and health problems, dust filters and activated carbon filters have been used for most space vehicles. Some

contaminants are dissolved into water that is then condensed on the heat exchangers, as well as the adsorbent used for CO_2 removal. The technology for removing trace contaminants in the water regeneration process is quite challenging. Thermal catalysts are used on the ISS to oxidize (mineralize) organics to CO_2 and water. This process effectively removes small organic molecules like methane and formaldehyde, which can be hard to remove by adsorption.

In order to reduce the energy requirements for the processing of trace contaminants, alternative concepts are being examined, such as photocatalytic removal by titanium oxide catalysts under UV radiation, and even biological filtration. If filters for removal of dust and particulates can be regenerated, it reduces the cost of life support. Novel approaches using electrostatic principles may alternatively be applied for this purpose.

17.2.2 Biological and Ecological Systems for Life Support

The use of biological and ecological systems in life support systems, ‘space agriculture’, is an advanced concept for future missions with large crews and longer durations. The ecological system, in general, is composed of three components: producer, consumer and decomposer organisms. Humans are the top consumer in agro-ecosystems. The main producers are photosynthetic organisms, such as plants, which convert the energy of sunlight to a chemical form that is fixed in biomass. Plants in space agriculture can also act as water distillers. Processed wastewater is irrigated, taken into the plant body, and transpired from leaves. Decomposers in the ecological system bridge the gap between consumers and producers, and drive materials recycling. In the space agro-ecosystem, human waste and inedible biomass could be composted by bacteria, and used to fertilize the soil or irrigation solutions with nutrients for farming more crop plants.

17.2.2.1 Water and Gas Regeneration by Space Agriculture

Supplying food, oxygen, and water for humans are the primary demands of engineering for living in space. If no element drops out of the recycling loop, the same amount of food should be regenerated from the wastes. The regeneration loop produces oxygen in a stoichiometric relationship to the biomass produced. When the scale of food production is large enough to provide more than 50 % of the crew’s diet, all the metabolic CO₂ produced by the crew is converted to O₂ in sufficient amounts to provide for the physiological needs of the crew. For water recycling, efficiencies may be gained by using the natural transpiration processes of plants. The plant transpiration ratio is that between the amount of water transpiration and photosynthetic fixation of energy (dry mass). The quantity of water that could be condensed from air in the farming area can be estimated from this ratio in ordinary crop plants. If food and oxygen produced by plants are to fill the requirements of the space crew, the amount of water recovered from the air in the system would typically exceed 200 L per person, which is close to daily consumption by a person on the ground. However, plant transpiration can be affected by the humidity and CO₂ concentration of the surrounding air, varying the evaporative gradient and the resistant to water flux from the leaves.

The degree of closure of materials recycling has long been the index for engineering a controlled ecological life support system (CELSS), also termed either a closed ecological life support system or a controlled environmental life support system. Space agriculture, like terrestrial agriculture, utilizes external resources available on the planetary body for its operation. Water, carbon dioxide, and bio-elements

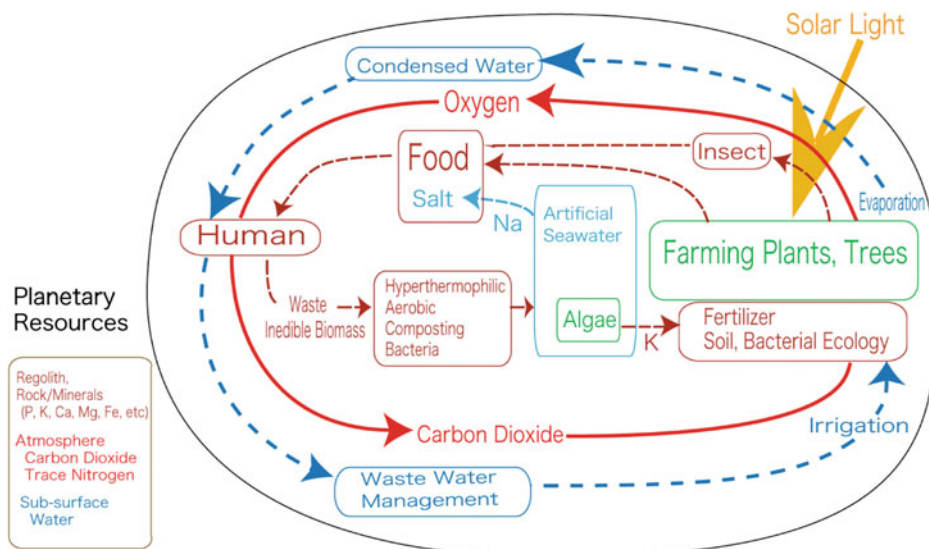
are the natural resources for agriculture. Gathering some of these in situ resources achieves more than 100 % closure, or in other words allows the disposal of some more recalcitrant components, while maintaining a materials balance. This enlarges the scale of the materials recycling loop from the initial inputs. This would be a form of in situ resource utilization (ISRU). As shown in Fig. 17.5, most of the required bio-elements are thought to be accessible on Mars. After space agriculture has been established and matured, systems might be expanded to even include trees to produce excess oxygen and sequester carbon in the form of wood, which could be used for living and habitat materials. Raising insects that feed on tree leaves or other cellulosic materials could further improve the efficiency of biomass use by providing protein rich foods.

17.2.2.2 Agro-Ecological Systems for Space Life Support

Selection of crop species is a fundamental step in designing space agriculture systems. Food materials and species should be selected to meet the nutritional requirements for healthy living, as described in Sect. 17.3.1. As the volume available for agricultural production in space is limited, and is a precious resource of space habitation, maximizing the yields of crops per unit area/volume per unit time is a high priority. The survivability of the crew will be greatly dependent on the robustness of crop growth and production. Toughness, including environmental resilience and pathogen resistance of agricultural plants, is also important for life support on a distant planet. Figure 17.6 shows one selection of food crops designed under these guidelines. This menu fills the nutritional needs for metabolic energy, dietary fiber, proteins and lipids (Fig. 17.7).

Considering the harsh conditions in space, a phased strategy to initiate agriculture is proposed. In order to develop this scenario, the trade-off between hydroponics and soil-based agriculture should be assessed during the various phases of development and integration of space agriculture. During the early phase, life support capability might be provided by a combination of the physicochemical systems and plant hydroponics, where the plants are only providing high-value supplemental or perishable foods. Hydroponics is robust because nutrient composition, dissolved oxygen level and other environmental factors for plant roots are strongly controlled. Composting of inedible biomass might then be used to gradually generate soils to support expanded agriculture. There are pros and cons for both hydroponics and soil-based farming. One advantage of using soil is the symbiosis between plants and microorganisms of the soil ecosystem, which often assist plant roots in obtaining essential elements to create a positive environment for plant growth. The soil ecology of bacteria, fungi, and other living organisms has been studied along with the physical and chemical

Fig. 17.5 Concept of space agriculture for habitation on Mars. If in situ resources are employed it is possible to make the system more than 100 % recyclable



properties of soil. Such knowledge provides a fundamental basis for initiating soil-based space agriculture [7].

The physiology of plants under micro- or partial gravity is another important subject to be studied for the engineering of space agriculture. Plants respond to gravity in two ways. They sense the gravity vector and develop their structures and orient their growth through gravitropism, also known as geotropism. They also respond to the magnitude of gravity by hardening cell walls to sustain their own body weight. Cellular mechanisms for both gravitropism and the gravi-resistance reaction are active research subjects in gravitational biology. In the absence of gravity, or under reduced gravity, tropisms induced by factors other than gravity, such as light and humidity, become more important. Since light is a governing factor for photosynthesis, it affects plant physiology in many ways. For some plant species, lighting and the length of day/night cycle should be controlled according to the photoperiodism that determines the timing of flower bud development and storage organ formation.

As with the cabin air, atmospheric pressure and composition in the farming section do not need to be the same as the air on Earth. As with the space habitat, lower pressures in farming modules will reduce the burden on the mechanical structure and reduce gas leakage. Limits of gas pressure and composition for farming plants have been examined in this context. The lower limit of oxygen for most plant species is probably near 10 kPa. An optimal range of carbon dioxide partial pressure is perhaps 100–200 Pa for C_3 plants.¹ Above this level, the photosynthetic reaction is saturated. On the other hand, transpiration water-use could increase [8]. And

of course human performance can be negatively affected at very high CO_2 partial pressures, which would have implications for humans tending the crops. Water vapor is another minor gas component that must be controlled for the transpiration of plants, because the water vapor pressure differential between the air and the leaves has a direct effect on transpiration rates. In a confined environment, the accumulation of bioactive substances must also be considered, if natural decomposition does not occur in the space agroecosystem. Ethylene gas is a particular concern, because it functions as a plant hormone that promotes flowering and ripening but can be harmful in higher concentrations. Ethylene can be removed by catalytic oxidation and possibly also through biological methods.

Heat and mass transfer phenomena change in exotic environments, such as reduced atmospheric pressure, different atmospheric composition, and micro- or partial gravity. The microclimate around plant bodies is modified under these conditions, and affects plant physiology in several ways. Natural convective heat and mass transfer driven by buoyancy is greatly suppressed under microgravity. This induces overheating of plant body parts, and suppresses gas exchange in the leaf. Diffusion of oxygen molecules is boosted under reduced pressure. These effects can be offset by maintaining adequate, forced air circulation throughout the plant production area. Oxygen, CO_2 , water vapor and other gas movements are increased at reduced pressure. A common observation from reduced pressure studies with plants is an increase in transpiration rates, although some plants may be able to adjust their transpiration rates as they acclimate to different pressures.

For the production of better quality entomophilous plant species, pollinator animals should be introduced. The flight capabilities of pollinator insects need to be confirmed under partial gravity and reduced atmospheric pressure, where both

¹ Together with C_4 carbon fixation and Crassulacean acid metabolism, also known as CAM photosynthesis, C_3 carbon fixation is a metabolic pathway for photosynthesis, converting carbon dioxide and ribulose biphosphate (RuBP, a 5-carbon sugar) into 3-phosphoglycerate.

Fig. 17.6 Model foods to satisfy the daily nutritional requirements for one person: rice (300 g), soybean (100 g), sweet potato (200 g), green-yellow vegetable (komatsuna) (300 g), silkworm pupae reared on mulberry leaves (50 g), loach fish co-cultured in rice paddies (120 g), and sodium salt (3 g)



Fig. 17.7 Cosmonaut Yuri I. Malenchenko, expedition 16 flight engineer representing Russia's Federal Space Agency, checks the progress of pea plants growing in the Russian Lada greenhouse in the Zvezda service module of the ISS. *Image NASA*



the fluid dynamics and lift force requirements differ from those on Earth. Adjustments of lift and thrust for flight maneuvers are made by commanding the muscles that control the attack angle or stroke trajectory of the wing motion.

Flight maneuvering of insects is controlled by visual sensory information. In parabolic flight experiments, bumblebees were able to handle the equivalent of gravity on Mars, but had difficulty under lunar conditions. Alternatively,

parthenocarpic or self-pollinating crops might be selected or developed for space applications. Besides pollination, raising animals may eventually become a necessary part of a space agriculture in order to fill the crew's nutritional needs for animal-origin food substances, as there are several nutrients that are difficult to obtain from plant-only diets; insects and fish are candidates for this supply.

Several species of insects have been proposed for the space diet. Silkworms, a well-established domesticated insect species, convert inedible mulberry leaves to edible materials; shown in Fig. 17.6. Among many fish species that can be bred, the tilapia and loach fish have been suggested. Co-culture of loach and rice is done in many places. Loach is a robust fish species. It gulps air into its digestive tube and expels it from its anus after exchanging oxygen and carbon dioxide through its gut. During the winter dry-up season when rice paddies lose their water, loach fish dive deep into the mud until spring. Furthermore, loach has a high nutritional value. A model diet with loach fish added to the selection of vegetables and insect could meet most nutritional requirements.

17.2.3 Waste Management

17.2.3.1 Solid Waste Management

To date, solid waste management has involved either consolidating or stabilizing solid waste and then dumping it to space, or returning it to Earth. This waste includes packaging materials, disposable clothes, food waste, and human metabolic wastes. Such wastes represent a potential source of odors and microbial pathogens. In order to reduce these risks, the waste can be dewatered and sealed with meltable plastics into inert disks or bricks. With the aim of recycling this waste, super-critical wet oxidation has been intensively studied. Organic substances are easily decomposed and oxidized by this method. However, the high power requirement and the handling of high pressures and temperatures are the major drawbacks of super-critical wet oxidation (Fig. 17.8).

17.2.3.2 Composting to Recycle Bio-Elements

Solid waste is a potential resource for space habitation and farming. The stabilized material might serve as a radiation shield, or produce mineralized compost for farming. This mineralization could be performed by physicochemical incineration with wet oxidation, or biological oxidation with microbes. One approach for partial recycling of minerals would be to treat solid wastes in aerobic or anaerobic stirred tank bioreactors, after which the nutrient rich effluent could be sent on to crops for food and O₂ production [9]. A more promising approach uses hyper-thermophilic aerobic composting. It is based on fermentation under high temperatures

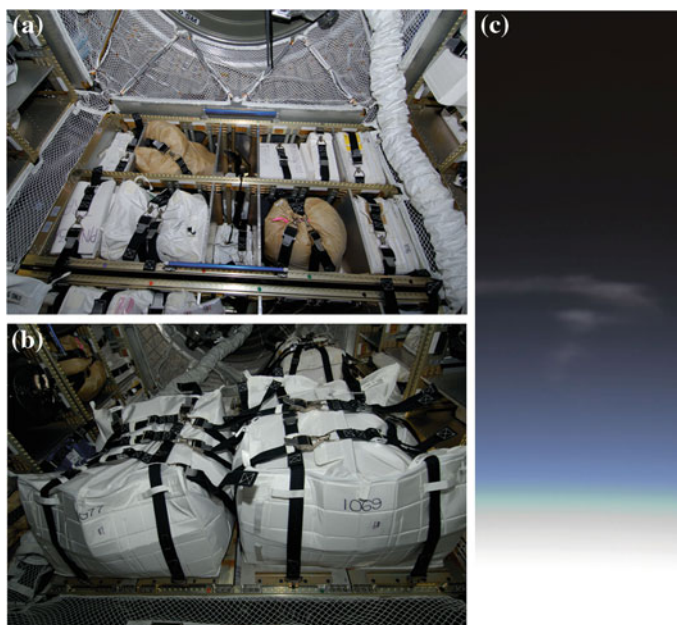
and aerobic conditions. This new composting system operates at a higher temperature than an ordinary anaerobic composting system. Bacteria in the hyper-thermophilic composter are active and viable at temperatures of 100 °C or even higher. This type of composting has a faster processing rate because of the rapid aerobic-type metabolic pathway, and a faster chemical reaction rate at the higher temperature.

Biological combustion releases heat and raises the temperature of the reaction bed, when air is force fed through it. Since microbial activity declines at higher temperatures (e. g., above 65–100 °C), the temperature in the reacting bed is naturally regulated to within its optimum range. Such natural regulation is effective if the volume of the reaction bed is large compared to its surface area. The ecology of this composting bacterial system is structured with an intensive symbiotic network formed among multiple species. There have been several terrestrial demonstrations of recycling essential nutrients (elements) from solid wastes back to plant production systems in the chemical forms accessible by plant roots. For this uptake, soil bacteria and arbuscular mycorrhizal fungi are important members of the symbiotic ecology in soil. Such an ecology is a primary advantage of soil-based agriculture.

For fertilization of crop plants, nitrogen in organic compounds is converted to ammonium ions through hyper-thermophilic aerobic fermentation. Organic nitrogen, typically amino or heterocyclic bio-chemicals, is either converted to ammonium or remains in an undigested form in the compost. The fate of phosphate in hyper-thermophilic aerobic composting has not yet been described in detail, although calcium phosphate precipitates have been found in reactors. The dropout rate of fertilizer elements in the recycling loop should be carefully examined to keep the degree of closure high. Potassium is the last of the three major macro-elements in fertilizer. Ring cavity structures formed in clay minerals can enclose potassium ions, which interact with the oxygen atoms surrounding the rim of the cavity. The high affinity of potassium ions for such sites, which originates in the interaction between the ions and oxygen atoms at the rim of the cavity, may function as a means of storing potassium to provide resistance to wash-out during watering.

Space agro-ecosystems may become saline as human wastes such as urine are increasingly processed in the materials recycling loop. This increased salinity can stress the crops and reduce productivity. Humans require sodium to maintain body fluid content, and any excess sodium is excreted in urine. Sodium should be separated, or reduced from the compost and fertilizer, in order to prevent a reduction of plant productivity. This can be a challenge while trying to retrieve nitrogen, potassium, phosphorus and other useful minerals from the waste stream. Several approaches have been proposed to solve this sodium

Fig. 17.8 The European Automated Transfer Vehicle (ATV), ISS supply spacecraft, is also used to dispose of solid waste. Shown are ATV-3, also known as ‘Edoardo Amaldi’, internal close-out photographs and it’s fiery plunge through Earth’s atmosphere and destructive reentry. *Image ESA (left) and NASA (right)*



problem in space agriculture. One approach would be to use physicochemical processes for separating the two elements, such as electro dialysis or struvite precipitation. The difference in temperature dependence on solubility of sodium and potassium salts could also be utilized to separate them by changing the temperature to drive the dissolution and precipitation cycle. Alternatively, biological processes might be able to partition sodium from potassium and other useful elements. One candidate technology for this purpose is the cultivation of marine algae to harvest potassium and other bio-elements from the medium, and increase the sodium content in artificial seawater. Another possible approach is the selection of salt tolerant halophytes as agricultural plant species.

17.2.3.3 Planetary Protection

When planning space missions to extraterrestrial bodies for astrobiological surveys, not contaminating them with terrestrial organisms or organic substances will have to be a mission priority. COSPAR, the international Committee on Space Research, has a review panel that assesses proposed mission plans under the consensus of space science. By this regulation, organic waste cannot be dumped from any spacecraft flying to Mars, or any other bodies that are of interest in terms of astrobiology. Hyper-thermophilic aerobic bacteria might be used to help to conserve the extraterrestrial environment for astrobiology exploration. The accumulated organic waste can be quickly composted with hyper-thermophilic aerobic bacteria, yielding a sterilized product. Eventually, EVAs will end up contaminating the surrounding areas of a surface outpost. For such future missions, the

target planets or bodies might be partitioned into safe or no-travel zones for humans in order to preserve their pristine nature. The issue of planetary protection will be discussed further in [Chap. 23](#), while [Table 17.4](#) details proposed categories for solar system bodies [10].

17.3 Life Management in Space

The daily life of the space crew requires more than just biological life support inside the cabin. Even when crewmembers are highly motivated to achieve the objectives of their mission, quality of life management enhances their productivity. Such management includes a wide range of items for achieving a fully balanced, healthy, and civilized lifestyle in space.

17.3.1 Food

In current missions to LEO space stations, food is stowed and resupplied from the ground at regular intervals. Fresh fruits and vegetables are delivered by cargo flights to the ISS, but consumed within days. Small-scale production of vegetables and other crops has been tested in space. However, larger scale crop-production systems for space life support will require solving complex integration challenges. The production of supplemental vegetables and other foods is the starting point for designing a space life support system [11]. The preferred selections of food materials are largely influenced by the backgrounds and food cultures of the crews.

17.3.1.1 Nutrition Requirements

Food provides a source of metabolic energy in the chemical form of molecules, fills the need for turnover of body, and supplements important substances for biochemical processes which cannot be synthesized in our body. Nutritional requirements are listed in Tables 17.5 and 17.6.

Humans are heterotrophic organisms and can oxidize carbohydrate, fat, or protein by oxygen to gain metabolic energy at a rate of about 100 W (roughly 2,000 kcal/day). This energy demand depends on body size, age, sex, and activity. Basal metabolic rate (BMR) is the synthesis of energy to maintain fundamental body functions. Maximum metabolic rate (MMR) is the BMR plus all neural and muscular work. Metabolic substrates (carbohydrates, fats and proteins) and pathways are used to meet physiological needs and condition.

Proteins and other constituents are required for the turnover of body tissue. The composition and amount of amino acids is important because an excess intake of protein results in the overload of urea-related metabolic waste in the liver and kidneys. Of the 20 amino acids needed for human nutrition, nine cannot be synthesized in the human body, and their intake from the diet is critical. The amino acid score is an index to evaluate the quality of amino acids composition.

Vitamins and other minor elements support health by supplementing nutrients that cannot be synthesized in our body. There are several general rules of thumb to guide the choice of foods to fill our nutritional requirements. One rule is the ratios of carbohydrate, protein, and fat necessary to meet the energy intake. Another is the recommended ratio of animal- and plant-origin proteins to optimize the amino acids score close to 100. Since carbohydrate and protein are hydrolyzed to smaller molecules of sugar and amino acid prior to their absorption through the intestinal cell membrane, the choice of animal- or plant-origin material is not essential for human nutrition. In many plant-origin foods, lysine is deficient. A combination of plant and animal substances solves this problem. In contrast to protein, animal- and plant-based fats and oils are different at their molecular level. Cholesterols are fatty acids found in animals but not in plants. A certain amount of cholesterols should be taken to keep the immune system functioning normally.

An example of how to improve crew health through food selection is to include foods with sufficient dietary fiber to keep the intestinal flora healthy. Also beneficial for this purpose are probiotics, which promote the ecology of intestinal microorganisms, such as populations of lactic acid bacteria. Space foods, at present, are sterilized and stabilized for long-term storage. For securing the shelf life of food, fermented foods like yogurt or pickles, which contain live bacteria, are not currently included in the space menu. Longer crewed missions should consider including probiotics in the menu, or feeding seed bacteria to establish healthy gut flora.

An additional consideration for nutrition specific to space crews relates to their intensive work associated with EVAs. Mineral loss from bone under micro- or partial gravity may be offset by taking supplemental Vitamin D, which also suggests that astronauts should be exposed to some minimal amount of UV light from time to time in order to sustain good Vitamin D nutrition. Although living in space is highly stressful in many aspects, the physical load during body movement is less under microgravity than on Earth. This factor should be taken into account when defining the nutritional requirements, and should avoid excess intake of energy.

17.3.1.2 Foods Storage and Cooking

Although fresh foods are supplied periodically to the ISS crew by cargo vehicles, most food is stored on-board for a long time. A shelf life of over one year at room temperature is the criterion for storage of food on the ISS. Microbial control measures are used in food preparation while trying to minimize any negative effects on the taste and texture. Foods can be stored either in metal cans or plastic packaging with low oxygen permeability. Oxidation accelerates food spoilage, and the chance of aerobic bacterial contamination is increased by the entry of permeated oxygen.

Thermo-stabilization, irradiation, and freeze-drying are common processes for preparing space food and for maintaining its quality throughout the mission period. Thermo-stabilization kills pathogenic and food-spoiling bacteria. Enzymes that impair the taste of food are also deactivated by high temperatures. Of the three processes, irradiation is a newer technology for food processing. Gamma-rays, X-rays, or electrons irradiate it at energies too low to produce radioactive isotopes. However, irradiated food is not fully accepted by the general public in many countries. The safety of irradiated food should be verified in terms of toxicity, oncogenesis, and the hereditary risk caused by probable reaction products of irradiation. Safety testing of irradiated foods after cooking or packaging is fairly new. Freeze-dried items and long shelf life bread have low water activity that is effective in suppressing microbial proliferation (Table 17.7).

Food and drinks are prepared on the ISS in its galley. Freeze-dried items are rehydrated in their plastic pouch with hot water, and canned or retort-packed food is warmed up in an electric oven similar to those that equip in the galley of a commercial passenger airplane. Cold beverages are also prepared using cold water supplied in the galley.

Dining is a highly social activity for humans. The menu should be structured with a variety of items and cooking methods to maximize palatability. For longer space missions, cooking will become one of the joys of space habitation [12]. Cooking facilities must meet safety regulations, and, moreover, fit into the space environment, such as micro- or partial gravity and reduced total pressure of cabin air.

Table 17.4 Proposed planetary protection categories for solar system bodies and types of missions

	Category I	Category II	Category III	Category IV	Category V
Type of mission	Any but Earth return	Any but Earth return	No direct contact (flyby, some orbiters)	Direct contact (lander, probe, some orbiters)	Earth return
Target body	Flyby, Orbiter, Lander: Venus; Moon; undifferentiated, metamorphosed asteroids; others TBD	Flyby, Orbiter, Lander: Comets; Carbonaceous Chondrite Asteroids; Jupiter; Saturn; Uranus; Neptune; Pluto/Charon; Kuiper-Belt objects; others TBD	Flyby, Orbiters: Mars; Europa; others TBD	Lander Missions: Mars; Europa; others TBD	Any Earth-return mission. 'Restricted Earth return': Mars; Europa; others TBD; 'Unrestricted Earth return': Moon; others TBD
Degree of concern	None	Record of planned impact probability and contamination control measures	Limit on impact probability Passive bioload control	Limit on probability of non-nominal impact Limit on bioload (active control)	If restricted Earth return: No impact on Earth or Moon; Returned hardware sterile; Containment of any sample
Representative range of requirements	None	Documentation only (all brief): Planetary protection plan Pre-launch report Post-launch report Post-encounter report End-of-mission report	Documentation (category II plus) Contamination control Organics inventory (as necessary) Implementing procedures such as: Trajectory biasing Cleanroom Bioload reduction (as necessary)	Documentation (category II plus) Probability of contamination analysis plan • Microbial assay plan Organics inventory Implementing procedures such as: Trajectory biasing Cleanroom Bioload reduction Partial sterilization of contacting hardware (as necessary) Bioshield Monitoring of bioload via bioassay	Outbound Same category as target body/outbound mission Inbound If restricted Earth return: documentation (category II plus) Pc analysis plan microbial reduction plan Microbial assay plan Trajectory biasing Sterile or contained returned hardware Continual monitoring of project activities Project advanced studies/research If unrestricted Earth return: • None

Inductive heating devices and pressurized pans are candidates for space cooking (Figs. 17.9, 17.10).

17.3.2 Clothes

Clothing choices are based on cultural tastes, crew comfort, and protection of the body from physical injury. Fabric

materials selected for the space crews' clothes cannot be toxic or flammable, and must meet safety criteria similar to those applied to other materials used inside the space cabin. Additional considerations are chemical stability, moisture absorption, water compatibility, strength, abrasion resistance, ease of cleaning, electrostatics, and freedom from lint. In designing clothes, the effects of microgravity on the body exposed to the space environment should be taken into

Table 17.5 Macronutrient guidelines for space flight [2]

Macronutrients	Daily dietary intake
Protein	0.8 g/kg and ≤ 35 % of the total daily energy intake and 2/3 of the amount in the form of animal protein and 1/3 in the form of vegetable protein
Carbohydrate	50–55 % of the total daily energy intake
Fat	25–35 % of the total daily energy intake
Ω -6 Fatty acids	14 g
Ω -3 Fatty acids	1.1–1.6 g
Saturated fat	<7 % of total calories
Trans fatty acids	<1 % of total calories
Cholesterol	<300 mg
Fiber	10–14 g/4,187 kJ

account. A person's height increases due to the spine lengthening in space. A shift in body fluid deforms (increases) the size of the chest and waist and decreases the size of the lower limbs. Neutral (resting) body posture differs from that under normal gravity, as shown in Fig. 17.11. The form of clothes is stabilized by the action of gravity on the ground. Clothing design in space should consider the absence of this effect. Altered heat and mass transfer under micro- or partial gravity is another factor to be considered when designing fabric.

The ISS has no laundry machine on-board, and the frequency of washing clothes is limited during a mission. The choice of surfactant for washing clothes should meet the constraints of the water revitalization system. Since most space clothes are disposable, the frequency of change determines the total quantity of clothes to be carried up. A daily change of underwear and workout clothes, and a weekly change of other clothing is the average frequency favored by crews. Functional fabrics and textiles are developed with a catalytic functional layer coated on the fibers to suppress odors.

17.3.3 Space Architecture

Space architecture is the practice of designing and building living environments for human activities in space. All items and subsystems that interface with the crew are the subject of space architecture. Its practice has two main components. One is to manage the unique environment of space and its constraints on habitation. The second is to find ways to humanize space systems and make them compatible with and user-friendly to crews.

Space cabin architecture needs to provide air circulation and illumination according to the requirements described earlier in this chapter. The layout and orientation of

Table 17.6 Micronutrient guidelines for space flight [2]

Vitamin or mineral	Daily dietary intake
Vitamin A	700–900 μ g
Vitamin D	25 μ g
Vitamin K	Women: 90 μ g, men: 120 μ g
Vitamin E	15 mg
Vitamin C	90 mg
Vitamin B12	2.4 μ g
Vitamin B6	1.7 mg
Thiamin	Women: 1.1 μ mol, men: 1.2 μ mol
Riboflavin	1.3 mg
Folate	400 μ g
Niacin	16 mg niacin equivalents
Biotin	30 μ g
Pantothenic acid	30 mg
Calcium	1,200–2,000 mg
Phosphorus	700 mg and $\leq 1.5 \times$ calcium intake
Magnesium	Women: 320 mg, men: 420 mg and ≤ 350 mg from supplements only
Sodium	1,500–2,300 mg
Potassium	4.7 g
Iron	8–10 mg
Copper	0.5–9 mg
Manganese	Women: 1.8 mg, men: 2.3 mg
Fluoride	Women: 3 mg, men: 4 mg
Zinc	11 mg
Selenium	55–400 μ g
Iodine	150 μ g
Chromium	35 μ g

components in the living and working sections should be consistent in terms of their vertical axis. A reversal of top and bottom between systems or areas induces confusion in the central nervous system, and causes space motion sickness. Providing visual cues to indicate vertical orientation, such as the contrast of light and dark colors for top and bottom, is effective in preventing such sickness. Figure 17.7 shows small portholes in the nadir face of the ISS that also serve to reinforce the sense of top and bottom.

For longer missions, psychology becomes an important factor in architectural design and in the choice of interior materials. Rather than plastics and metals, natural materials such as wood may be preferable for those items with which the crew directly interface. However, all proposed materials and architectural design might be subjected to the safety control of space systems. Escape routes and passages should be securely cleared in case of fire or other catastrophic events.

The crew quarters are a compartment for sleep and privacy, although minimally sized. Windows weaken

Table 17.7 Types of food and packaging for space flight [2]

Food/packaging type	ISS/Space shuttle example	Parameters
Thermostabilized	Beef stew	Shelf life: 3–5 years Packaging: quad-laminate pouch Preparation: none or heating
	Yogurt	
	Pudding	
	Soup	
	Tuna casserole Red beans and rice	
Irradiated	Beef brisket, fajitas	Shelf life: 2–5 years Packaging: quad-laminate pouch Preparation: none or heating
	Broiled lamb	
	Fresh fruit	
	Raw vegetables	
Rehydratable vegetables	Chicken salad	Shelf life: 1.5 years with overwrap; 1 year with no overwrap Packaging: combitherm pouch, adapter for rehydration Preparation: rehydration using hot water
	Cornbread dressing	
	Sausage patty	
	Shrimp cocktail	
Natural form cookies	Brownies	Shelf life: 1.5 years with overwrap; 1 year with no overwrap Packaging: combitherm pouch Preparation: none
	Nuts	
	Granola bars	
Extended-shelf-life bread products	Dinner rolls	Shelf life: 1 year Packaging: Preparation: None
	Waffles	
	Scones	
	Tortillas	
Fresh food	Fresh fruit	Shelf life: 1 week Packaging: Preparation: None
	Raw vegetables	
	Tortillas	
Beverages	Freeze-dried (coffee or tea)	Shelf life: 1.5 years Packaging: tri-laminate pouch, adapter for rehydration, straw Preparation: rehydration using hot or cold water
	Drink mix (lemonade)	
	Water	

pressurized modules by producing non-uniformity in their mechanical structure. However, this is counterbalanced by the psychological benefit of being able to view outer space and the home planet. Among space crews, rank hierarchies and leader–follower relationships are formed. Space architecture can reinforce the stability of this group structure, by having the layout of crew quarters and other items ordered by rank (Figs. 17.12, 17.13).

Hygiene and other items of daily life are managed by the space architecture design as well. Space toilets separate urine and feces for water regeneration and storage prior to dumping or composting. These are facilities and tools for shaving, hair cutting, tooth brushing, hand washing, and bathing. All of these capabilities must be integrated with solid and liquid waste handling systems, either for stabilization and storage, or for recycling and loop closure for future missions.

17.3.4 Medical Support

A major question asked during the early era of space flight was whether humans could sustain life in outer space without major health problems. This issue is not yet fully resolved. Of the many factors in the space environment, research has focused on gravity and its compounding effects with other factors such as space radiation. In space, gravity can be handled as an experimental parameter for periods long enough to produce biological effects. Experimental results to date indicate that microgravity does not cause genetic instability. Extensive studies have been conducted to clarify the direct or secondary effects of gravity. Its actions have been surveyed in the several hierarchical levels of living systems, i.e., cell, organ and whole body. Based on those space-specific findings, medical support is provided for astronauts.

Fig. 17.9 Astronauts Shane Kimbrough and Sandra Magnus, both STS-126 mission specialists, are pictured with fresh fruit floating freely on the middeck of Space Shuttle Endeavour during flight day three activities. *Image* NASA; S126-E-007618 (16 Nov. 2008)



Fig. 17.10 Japan Aerospace Exploration Agency (JAXA) astronaut Koichi Wakata (若田 光一), expedition 18/19 flight engineer, is pictured with food and drink containers floating freely in the Harmony node of the International Space Station. *Image* NASA; ISS018-E-044614 (4 April 2009)

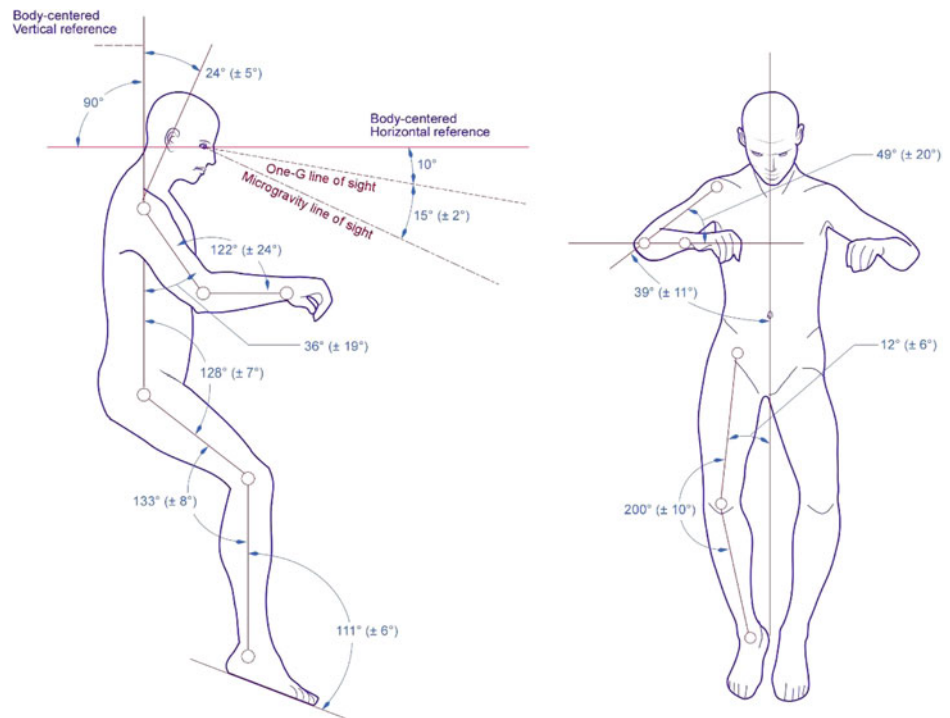


17.3.4.1 Human Physiology in Space

Through the evolutionary history of living organisms on Earth, the human body and its physiological features have become well adapted to the planetary environment. Once the human body is exposed to microgravity, adaptation to the new environment starts. Neurophysiology related to graviperception changes after a transition to microgravity. Sensory inputs from vestibular organs, eyes, and somatic senses are integrated in the brain to recognize body orientation and control posture. Space motion sickness (SMS) is often reported during the first few days in space, and is thought to

be induced by confusion in this sensory integration process. Body fluids shift from the lower limbs to the upper body under microgravity. The cardiovascular system adapts to the space environment too. Motor function changes in response to the absence of mechanical load on the muscles that would otherwise be counteracting gravity. The muscle fibers are gradually modified to fast-twitch type muscle fiber, and overall muscle tissue atrophies. Continuous bone demineralization is a critical risk of longer duration space flights. Each of these adaptations happens over a different time span, and some are irreversible even after return to Earth.

Fig. 17.11 Neutral body posture under microgravity [2]



The action of gravity on early developmental processes has been carefully examined with several animal models, but is not yet completely understood.

17.3.4.2 Health Monitoring and Medication

Crew health status is monitored both on-board and remotely from the ground at a certain interval during normal operation, and continuously during critical operation. Preventive health care is conducted by diagnosis of routine exams, such as blood pressure and electrocardiogram (ECG). Crewmembers are periodically interviewed by a flight surgeon on the ground for consultation on medical and psychological issues.

The crew health care system (CHeCS) on-board the ISS consists of the health maintenance system (HMS), environmental health system (EHS), and countermeasures system (CMS). Blood samples are analyzed with a clinical chemistry kit. Saliva samples are also collected for analysis.

Pharmacological countermeasures to SMS involve a drug that selectively modifies the related neural activities. To prevent bone mineral loss, an effective drug is prescribed for space crew. In addition, physical exercise can prevent some muscle atrophy, if it is caused mainly by disuse, and there are no major direct effects of gravity on cellular processes. The crew is provided with physical exercise protocols and equipment such as a treadmill, cycle ergometer, and resistive exercise device. By applying compression force to skeletal bones, demineralization or remodeling is suppressed at a certain level. Orthostatic intolerance after return to Earth is

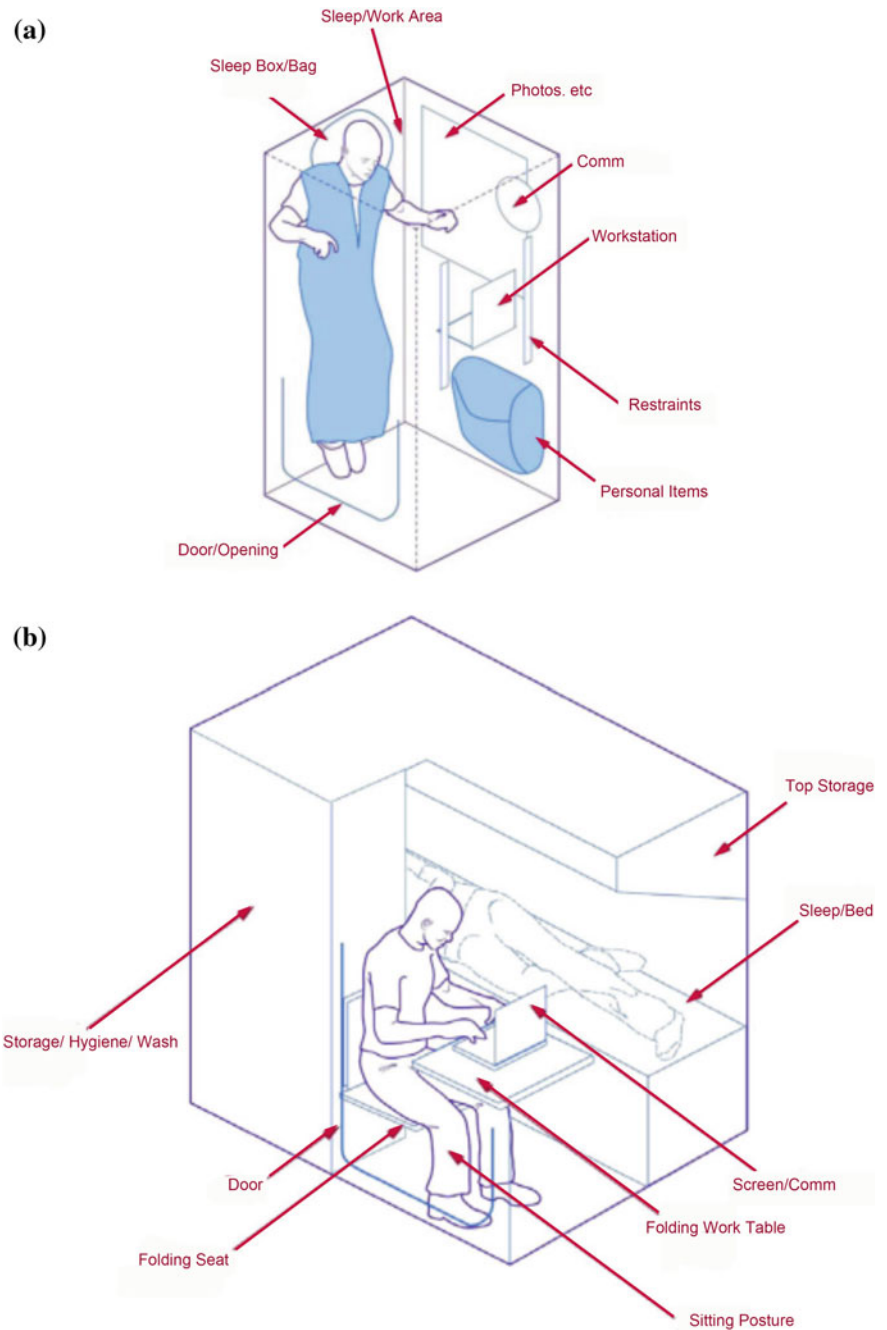
still an unsolved problem in space medicine. Lower body negative pressure (LBNP) and oral fluid loading is prescribed to increase total body fluid, just prior to return to Earth. As noted earlier, long-term exposure to space radiation is still a challenge for crew health, and medications for countering radiation effects are still in the research phase.

The medical facility for therapeutic operation provides the capability to administer hypodermic and intravenous injection. Medical risks that cannot be well evaluated before launch are ranked by severity. Two high risks, which require surgery without delay, are appendicitis and heavy or traumatic injury. A crew medical restraint system (CMRS) is prepared on-board, together with surgical tools. A defibrillator is also available for heart attacks. The handling of a deceased crewmember's body is another integrated function that must be considered, especially for long-duration missions. It is not realistic to include a medically licensed crewmember on the team in all missions, and even medical personnel find it difficult to cover every field of medicine. Telemedicine has been identified as a key technology for such situations. Medical experts can help to diagnose conditions and prescribe medication remotely, thereby easing the burden on the crew.

17.3.5 Human Factors and Psychological Support for Space Exploration

The crew is the core component of any human space system, but human factors are the most difficult to define in systems

Fig. 17.12 Design of sleep areas and crew quarters. **a** Crew quarters in microgravity. Minimal room for sleeping, working, and rest, which are all done in the same space for short to medium duration missions. **b** Partial-gravity crew quarters. Large volume for long-duration missions. Working, sitting, standing, and sleeping combined with hygiene, stowage, and waste management



engineering. Human factors should be properly handled in order to ensure mission success. Social psychology is a dominant factor in the success of prolonged duration space missions. In addition to the harsh physical factors of the space environment, psychological stressors for astronauts are isolation from their home, absolute confinement in a narrow cabin surrounded by the vacuum of space, the danger of the mission as a whole, the heavy workload, and the monotony of life during a mission. Personal behavior and productivity are definitely influenced by an astronaut's psychological state.

Crew team performance is largely affected by interpersonal relations among crewmembers, and group dynamics are key to maximizing the morale and productivity of a crew. Group structure, the composition of crew, and leader/follower relationships are important subjects to be examined. In the short term, heterogeneity among team members, such as gender, career experience, language and cultural background, does not give positive effects. After the acceptance of differences by each crewmember, however, the diversity among space crew is a positive feature in interpersonal relations. Gender of space crew has been treated as a

Fig. 17.13 Visual access through a window. Being able to view outside the spacecraft or space habitat is important for crew psychology. Japanese Aerospace Exploration Agency (JAXA) astronaut Koichi Wakata (□□ □□), expedition 19/20 flight engineer, takes pictures out the forward Kibo or JEM window on the International Space Station. *Image* NASA; ISS020-E-021689 (17 July 2009)



sensitive issue, and is certainly important. Cultural differences associated with gender can be understood in the context of the evolution of social behavior of humans. Depending on the cultural background, women in a mixed-gender group can take on a socio-emotional role, instead of adopting a highly competitive attitude in a task-oriented team. Between same aged members of both sexes, rivalry and harassment are likely to be serious problems under stressful conditions [13]. Nevertheless, in all cases there can be substantial variation among individuals.

Psychological support for crews who are confined and isolated in a spaceship becomes an essential part of life management for interplanetary travel, which is characterized by a longer mission period, extreme isolation, and physical distance from home. Broad studies in the field of behavioral and biological sciences are required to understand and solve these problems. In order to maximize both the motivation of crewmembers and the performance of the group as a whole, extra care should be taken with respect to their psychology.

For this purpose, a recreational facility might be important, and should be large enough to accommodate all crewmembers at the same time. Games are an excellent recreation, and promote interpersonal interactions in a meta-world. Setting handicaps among crew members avoids fixed winners and losers of games. Dining is a cultural and social event. Celebration dinners can include an appropriate amount of alcoholic beverage at certain intervals, depending on the crewmembers cultural backgrounds. Decision making for dining among crewmembers may help to maintain the leader and follower relationships. Keeping green plants and live animals as pets in the spaceship or outpost will also have positive effects on crew psychology and interpersonal relations. A core part of the psychological support for a space

crew is communication with ground personnel, including medical and psychological counselors, as well as family and friends.

17.3.6 Innovative Approach for Human Space Mission Challenges for the Final Frontier

If it is decided to extend human space exploration beyond the Moon, an innovative approach will be required for life support management. For larger crew sizes and longer mission durations, the bioregenerative life support concept should be given strong consideration for recycling bio-substances. However, there are many issues to be solved before its implementation. The robustness of space agriculture must be verified to ensure survivability of the space crew, and the risks of system failures must be understood. Human factors and psychology, together with a broad range of other new items, are essential fields of study prior to making crewed missions to Mars and other distant extraterrestrial bodies the final frontier of human civilization.

References

1. Tsiolkovsky, K.E. 1926. Exploration of world space with rockets. Kaluga Press, Russia.
2. NASA/SP-2010-3407, Human Integration Design Handbook (HIDH) January 27, 2010.
3. Law J, Watkins S, Alexander D; In-Flight Carbon Dioxide Exposures and Related Symptoms: Association, Susceptibility, and Operational Implications, NASA/TP-2010-216126, (2010).

4. NASA/JSC Toxicology Group, "Spacecraft Maximum Allowable Concentrations for Airborne Contaminants", JSC 20584, June 1999.
5. SSP 50005C, International Space Station Program Flight Crew Integration Standard, NASA-STD-3000/T, 15 December 1999.
6. "NASA Facts", FS-2002-05-85-MSFC, NASA MSFC, May 2002.
7. Doug Ming and Don Henninger (eds.); Lunar Base Agriculture: Soils for Plant Growth. Amer. Soc. Agronomy, Madison, WI, USA. 255 pages (1989).
8. Wheeler, R.M., C.L. Mackowiak, N.C. Yorio, and J.C. Sager. 1999. Effects of CO₂ on stomatal conductance: Do stomata open at very high CO₂ concentrations? *Annals of Botany* 83:243–251.
9. Mackowiak, C.L., R.M. Wheeler, G.W. Stutte, N.C. Yorio, and J.C. Sager. 1997. Use of biological reclaimed minerals for continuous hydroponic potato production in a CELSS. *Adv. Space Res.* 20 (10):1815–1820.
10. J.D. Rummel, P.D. Stabekis, D.L. Devincenzi, J.B. Barengoltz; COSPAR's planetary protection policy: A consolidated draft. *Advances in Space Research*, 30, 1567–1571 (2002).
11. MacElroy, R.D., M. Kliss, and C. Straight. 1992. Life support systems for Mars transit. *Adv. Space Res.* 12(5):159–166.
12. H.W. Lane and D.A. Schoeller (eds.) Nutrition in space flight and weightlessness models. CRC Press, Boca Raton, FL, USA. pp. 41–67.
13. Nick Kanas, Dietrich Manzey; Space Psychology and Psychiatry, Springer (2010).
14. P. Spillantini, M. Casolino, M. Durante, R. Mueller-Mellin, G. Reitz, L. Rossi, V. Shurshakov, M. Sorbi; Shielding from cosmic radiation for interplanetary missions: Active and passive methods. *Radiation Measurements*. 42, 14–23 (2007).
15. Committee on Solar and Space Physics and Committee on Solar-Terrestrial Research, National Research Council: Radiation and the International Space Station: Recommendations to Reduce Risk. National Academy Press (2000).
16. Harry W. Jones, Mark H. Kliss; Exploration Life Support Technology Challenges for the Crew Exploration Vehicle and Future Human Missions. *Advances in Space Research*, 45, 917–928 (2010).
17. Jay C. Buckley; Space Physiology. Oxford University Press, (2006).
18. Masamichi Yamashita, Hirofumi Hashimoto, Hidenori Wada; On-Site Resources Availability for Space Agriculture on Mars, in MARS: Prospective Energy and Material Resources, Edited by Viorel Badescu, Springer-Verlag (2009).
19. Raymond M. Wheeler; Plants for Life Support: From Myers to Mars, *Gravitational and Space Biology*, 23, 25–34 (2010).
20. Iosef I. Gitelson, Genry M. Lisovsky; Creation of Closed Ecological Life Support Systems: Results, Critical Problems and Potentials. *Journal of Siberian Federal University. Biology*, 1, 19–39 (2008).
21. Nelson, M and W.F.Dempster,1996, Living In Space: results from Biosphere 2's initial closure, an early testbed for closed ecological systems on Mars,pp.363-390 in Strategies for Mars: a guide to human exploration ed.C.R.Stoker & C.Emment, Vol.86 AAS Publication, San Diego CA.
22. Nelson, M., N.S. Pechurkin, J.P. Allen, L.A. Somova and J.I. Gitelson, 2010. Bioengineering of Closed Ecological Systems for Ecological Research, Space Life Support and the Science of Biospherics, Chap. 11 in Volume 10, 2010, doi: [10.1007/978-1-60327-140-0](https://doi.org/10.1007/978-1-60327-140-0) ENVIRONMENTAL BIOTECHNOLOGY in the Handbook of Environmental Engineering series, Editors: Lawrence K. Wang, Volodymyr Ivanov and Joo-Hwa Tay, The Humana Press, Inc., Totowa, NJ.
23. Keiji Nitta; The Mini-Earth Facility and Present Status of Habitation Experiment Program. *Advances in Space Research*, 35, 1531-1538 (2005).
24. A. Scott Howe, Brent Sherwood; Out of This World: The New Field of Space Architecture, Ned Allen, Library of Flight, AIAA, (2009).
25. Jack W. Stuster; Bold Endeavors: Behavioral Lessons from Polar and Space Exploration. *Gravitational and Space Biology Bulletin* 13(2), 49–58 (June 2000).

Further Reading

Steve Lingard and John Underwood

18.1 Overview

18.1.1 Definitions

The process of delivering a payload from an interplanetary transfer trajectory or from a planetary orbit to a stationary position on the ground may generally be split into three phases: entry, descent, and landing.

Entry is the part of the trajectory from first contact with the destination's atmosphere until either additional aerodynamic deceleration is deployed or the landing phase commences.

Descent is the part of the trajectory characterised by the action of one or more aerodynamic decelerators which reduce the velocity to a value compatible with the landing system.

Landing is the final deceleration to come to rest on the surface of the planet or moon.

18.1.2 Entry

For a body with an atmosphere, the entry phase is responsible for dissipating the vast majority of the kinetic energy arising from the interplanetary trajectory cruise or from a planetary orbit. A typical relative velocity at the start of entry is between 4 km/s for Mars orbit and 47.4 km/s for direct entry to the Jovian atmosphere from interplanetary cruise [1]. Aerodynamic decelerators such as parachutes or ballutes only function correctly at much lower velocities. The excess kinetic energy must therefore be dissipated by the entry system.

The excess kinetic energy is generally converted to heat, by means of the aerothermodynamics of the entry vehicle.

This energy is then dissipated by a combination of radiation, convection, and endothermic chemical reactions.

The configuration of the entry vehicle during the entry phase is generally unchanged from the cruise phase, other than potentially the jettisoning of a cruise instrument unit. The shape of the vehicle is designed to facilitate the controlled dissipation of the kinetic energy by means of its aerodynamic shape and mass properties. A recent exception to this case has been with the development of inflatable aeroshells to increase the drag area of the vehicle and thus increase the deceleration for given flow conditions.

The entry phase is generally unguided. The entry vehicle is trimmed to fly in a non-lifting configuration. This is the simplest form of trajectory and is suitable for most planetary entries with sufficiently dense atmospheres and robust payloads.

Lifting trajectories may be achieved by offsetting the center of gravity of the vehicle so that it flies at a constant, non-zero angle of attack. This allows limited control of the flight path angle and cross-range velocity to be achieved by rotating the vehicle and thus the lift vector about its axis. Lifting trajectories allow limited targeting and the moderation of accelerations and heating by continuously targeting an appropriate altitude and thus density for the desired parameters.

Finally, the vehicle can utilize a fully controlled entry such as that of the Space Shuttle orbiter or X-37 spaceplane. This allows far better control of the entry than a simple lifting vehicle but also requires significant control input.

18.1.3 Descent

The terminal velocity that may be achieved during the entry phase is limited by the area of the entry vehicle, which is in turn usually limited by the size of the fairing on the rocket used to launch it. This velocity is usually too large for any practical landing system. Thus, the velocity must be reduced still further during the descent phase.

S. Lingard (✉) · J. Underwood
Vorticity Ltd., Chalgrove, UK
e-mail: Steve.Lingard@vorticity-systems.com

The prime purpose of the descent system is to reduce the descent velocity to a value compatible with the landing system. Two general types of descent system are in common use: aerodynamic decelerators and powered descent. The first type usually consists of one or more aerodynamic decelerators; usually parachutes. These also fulfill secondary roles such as stabilization of the vehicle during descent and removal of the protective aeroshell from the lander. The second uses a rocket system to reduce the vehicle velocity and usually guide it towards a safe landing site. Hybrid systems are also possible in which a parachute performs the initial deceleration and the descent sequence is completed using powered descent.

Parachute descent systems can consist of anything from one parachute upwards. Multiple parachute systems may be multi-phase (where the parachutes are deployed sequentially), clustered (where several parachutes are deployed at the same time), or both.

Multi-phase systems are usually chosen where there are staging requirements (e.g. removal of a two-part aeroshell) or where the use of multiple parachutes allows mass reduction of the system by optimization of the parachutes for their operation conditions. In a single parachute system, a parachute type that is compatible with supersonic deployment and stable terminal descent must usually be chosen. It must be sized for the required terminal velocity but stressed for the supersonic inflation force. Use of a multi-parachute system allows a small, robust, supersonic-optimised parachute to decelerate the payload to a low dynamic pressure where a large, lightly stressed, stable parachute can complete the deceleration to landing.

Clustered parachutes may be chosen either to improve the control of stresses or, more usually, to provide redundancy in case of failure of one canopy. They also allow the use of otherwise unstable parachutes, since a cluster will always descend vertically with respect to the air mass.

Another example of an aerodynamic decelerator system is a lifting body such as the Space Shuttle orbiter. In this case, the descent system allows the vehicle to be flown as a glider to reach a conventional runway.

Powered descent systems are chosen where soft touchdown or terminal guidance is required, or where there is insufficient atmosphere for an aerodynamic decelerator system. They require multi-nozzle or multi-engine, throttleable rocket systems, an inertial guidance platform, and a guidance system.

18.1.4 Landing

The landing system completes the process of bringing the vehicle to rest on the surface of the planet by absorbing the remaining kinetic energy from the descent phase. This can

take many different forms. For a water landing, the landing energy may be absorbed by displacement of the liquid in which it lands. For a landing on a solid surface, the energy may be absorbed using crushable materials, landing legs, or airbag systems. In some cases, a proportion of the final energy may be removed just before landing by a retro-rocket system. The choice of system depends largely on the mass of the vehicle, the mission requirements, and the terminal velocity.

For low terminal rates of descent, such as those resulting from a powered descent or an aircraft-type landing, either a crushable material or landing legs may be used. These are mass-effective where very small strokes are required but they are limited for larger rates of descent by the need for longer strokes to limit accelerations to acceptable levels.

Airbag systems are capable of accommodating larger terminal rates of descent. The size of the airbag system is determined by the terminal rate of descent, acceptable vehicle acceleration, and thermodynamic efficiency of the system. Airbag systems work by converting the kinetic energy of the vehicle into internal energy of the gas within the airbags. Two types of airbag system have been used: vented and non-vented, the difference being in the way in which they manage the energy absorbed during the landing event.

A vented airbag system detects when the payload has, or is about to, come to rest. It then activates a valve that allows the compressed gas within the airbag to vent to the atmosphere, thus releasing the energy. This results in the payload coming to rest on the first impact and allows the airbag system to be designed to protect only the lower side of the vehicle. The terminal trajectory of the payload must be considered carefully in order to preclude the vehicle tipping over during the landing event.

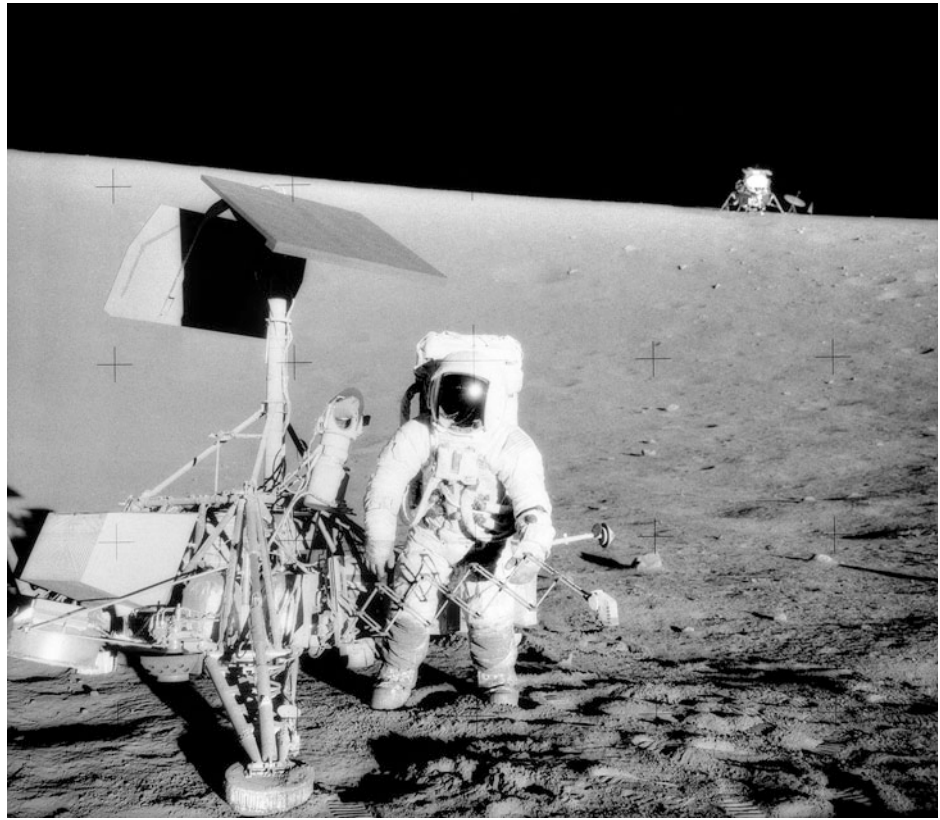
In a non-vented airbag system, the pressure within the airbag is returned to kinetic energy and the vehicle bounces. This will then result in a series of subsequent impacts. In this case, the initial impact energy is gradually dissipated over a number of bounces by means of thermodynamic losses in the compression and expansion of the airbag gas and by deformation of the surface on which it lands. Non-vented airbag systems are simple, requiring no control systems; however, they must envelop the lander, since the orientation of the second impact is random, and the airbags must subsequently be released or retracted in order to enable the lander to perform its function.

18.1.5 Examples

The range of potential entry, descent, and landing systems (EDLS) is best illustrated using mission examples.

The lunar Surveyor landers were typical of landing systems for bodies without atmospheres. The entire

Fig. 18.1 Charles Conrad Jr., Apollo 12 Commander, examines the robotic Surveyor III spacecraft during the second extravehicular activity (EVA-2). The Lunar Module (LM) 'Intrepid' is in the right background. This picture was taken by astronaut Alan L. Bean, Lunar Module pilot. The Intrepid landed on the Moon's Ocean of Storms less than 200 m from Surveyor III. The television camera and several other components were taken from Surveyor III and brought back to Earth for analysis. Surveyor III soft-landed on the Moon on April 19, 1967. *Image* NASA; AS12-48-7136 (20 Nov. 1969)



deceleration from transfer orbit to landing was achieved using a powered descent, decelerating from 2 km/s to a landing velocity of only 3.5 m/s. The landing system consisted of three articulated legs that included energy absorbing elements to remove the final descent velocity [2], as can be seen in Fig. 18.1.

The Mars Pathfinder lander was designed to be a simple, reliable entry, descent, and landing system. The entry phase was accomplished using a non-lifting ablative aeroshell that decelerated the lander to Mach 1.8, at which time a single parachute was deployed. This decelerated the vehicle to a velocity of 63 m/s, at a height of 88 m above the surface, where upon a retro-rocket system decelerated it to rest at a height of 21 m and a non-vented airbag system was deployed [3], as shown in Fig. 18.2. The lander came to rest following a series of bounces.

The Apollo command module used a lifting, ablating aeroshell for the entry phase [4]. This allowed management of the heat load, deceleration, and landing point. The descent system comprised of eight parachutes: two drogues for initial stabilization; and three pilot chutes, each of which deployed one of the three main chutes [5]. The system was sized to survive single parachute failures. A water landing was chosen so that no landing system was required other than post-landing flotation devices.

The Space Shuttle orbiter was a lifting body with an insulating coating. The entry was controlled to keep the heating rates within a range that could be dissipated during entry. The descent and landing was accomplished as a glider.

18.2 System Design and Design Drivers

When designing a planetary lander or Earth return vehicle, it is important that the entry, descent, and landing elements be designed as an integrated system rather than as three separate entities. It is also important that the EDL system design is considered during the initial mission design when considering approach trajectories, landing sites, and direct versus orbital entry.

The system design drivers may be split into two groups: those that are outside the control of the mission designer (e.g. the atmosphere and gravity of the destination) and those that may be influenced by the EDL designer (e.g. the approach velocity and system mass).

The atmosphere of the target body is a key driver in the design of an EDL system. If the body has no atmosphere, then a powered descent or a hard landing are the only options for the entry and descent. For a body with an

Fig. 18.2 Engineers test the Mars Pathfinder airbags in June 1995



atmosphere, the nature of the atmosphere is important to the system design.

The atmosphere has a direct influence on the available entry conditions. A steep atmospheric entry allows accurate targeting, albeit with high deceleration and heat flux. For a dense atmosphere such as that of Earth, the steepest entry is limited by the allowable heat flux and deceleration. For a tenuous atmosphere such as Mars, the limit is determined by the need to complete the entry and descent deceleration before reaching the ground. A shallow entry reduces both acceleration and heat flux, but risks skipping out of the atmosphere if too shallow. For tenuous atmospheres such as Mars, the useable entry corridor for an arrival from interplanetary cruise is only a few degrees wide.

Knowledge of atmospheric properties and their variability is a second important influence. The entry and descent system must be compatible with the likely range of atmosphere properties. Where the atmosphere properties are not known or are known but are variable, margin must be taken in system design to allow for this.

The gravity of the target body, along with the atmosphere, determines the potential solutions for terminal descent. Where a parachute system is used, gravity will influence the size of parachute required for a given rate of descent such that either a powered descent or a retro-assisted terminal descent may be appropriate.

The entry velocity is determined by the gravity of the body and the arrival conditions (e.g. from orbit or interplanetary cruise). An entry from orbit will have a lower energy than a direct entry; however, this is usually at the expense of an inefficient chemical burn to place the spacecraft in orbit initially. In practice, direct entries are used for all planetary missions unless there is a compelling

reason for a pre-entry orbital phase (for example, in the case of Viking when the martian landing sites were assessed before initiating the lander mission).

The ballistic coefficient, introduced in [Chap. 5](#), ($\beta = m/C_d S$) of an object is a measure of its terminal velocity or resistance to deceleration. A low ballistic coefficient is advantageous for entries to planets with thin atmospheres since the deceleration to landing velocity may be achieved more rapidly. The minimum ballistic coefficient is generally limited by the launch vehicle. That is, because heat shields are usually solid, their maximum diameter is limited to that of the launch fairing. This results in a maximum mass of entry vehicle that can be considered for a Mars entry without using novel techniques such as inflatable aeroshells.

Throughout the design process, the ability to qualify the EDL system must be considered. The best technical solution is not always the most practical to validate. The needs of validation must always be considered against those of technical elegance.

18.3 Entry

The objective of the entry phase is to reduce the relative velocity of the entry vehicle to an appropriate value for the transition to the descent/landing system at a point where it has sufficient altitude and time to operate. This is accomplished by converting the kinetic energy of the vehicle into thermal energy by means of the aerothermodynamics of the vehicle and then either dissipating it (by ablation, convection, or radiation) or absorbing it. The system charged with preventing the heat generated during entry from damaging the entry vehicle is the thermal protection system (TPS).

18.3.1 Concepts and Constraints

For any mission, there will be an acceptable entry corridor. The flight path angle at the start of entry is limited at the shallow extreme by the risk of skipping out of the atmosphere and at the steep extreme by lack of altitude, excessive heat flux, or deceleration.

For a shallow entry angle, there is the possibility of skipping out of the atmosphere if the velocity is high enough, rather than entering as intended. During the initial phases of entry, the flight path angle will reduce (become shallower) as the surface of the planet falls away from the vehicle. The deceleration of the vehicle must be sufficient to allow it to follow the curvature of the planet. The risk of skipping is greatest for small bodies (with the smallest radius), tenuous atmospheres, and shallow entry angles.

For a given entry velocity, the kinetic energy to be dissipated during the entry is fixed. The designer does, however, have control over both the magnitude and the length of the heating event, and this allows the thermal protection system to be optimized.

The peak heat flux occurs part way through the entry. The heat flux increases with increasing atmospheric density but decreases with reducing velocity. For a given entry vehicle and arrival velocity, the peak heat flux will be greatest for a steep entry where the vehicle penetrates the denser atmosphere while retaining more of its initial velocity. The peak heat flux determines the type of ablative material that may be used in the heat shield.

The peak acceleration (i.e. deceleration) usually occurs close to the peak heat flux. Again, the greatest acceleration is generally associated with the steepest entry. The maximum allowable acceleration may be limited by the design of the entry vehicle or by the physiological limits of its occupants.

The total heat load into the thermal protection system depends on the characteristics of the entry trajectory. While the total energy dissipated during an entry is fixed for a given entry velocity, the proportion which is absorbed rather than radiated or convected away depends on the characteristics of the entry vehicle and its trajectory.

The heat soak must be considered during the design of a protective heat shield. During entry, much of the kinetic energy will be dissipated; however, the surface temperature of the thermal protection system will be elevated and heat will be conducted to the vehicle behind it. The TPS must be designed to prevent the protected structure from reaching unacceptable temperatures. For planetary entry vehicles, it is conventional to design the TPS such that the heat pulse does not reach the rear face of the TPS until after the heat shield has been released. This leads to a requirement to reduce the entry trajectory so that there is less time for the heat to soak through the TPS. An alternative, employed on

the Space Shuttle orbiter is to absorb the heat and re-radiate it before it has time to soak through to the vehicle beneath.

During entry, it is usually necessary for the vehicle to remain in a stable attitude in order for the TPS to work optimally. This requires either an active control system or, more usually, a vehicle that is aerodynamically stable over the deceleration from hypersonic to supersonic speeds. This places constraints on the shape of the entry vehicle and on the location of the center of gravity.

18.3.2 Entry Heating

During entry, the kinetic energy of the vehicle is converted to heat energy. This principally occurs in the shock wave ahead of the vehicle. For this reason, entry vehicles are designed with blunt noses to ensure that the shock wave stands off from the vehicle. The transmission of the energy to the vehicle is by both convection and radiation. As the atmospheric gas passes through the shock wave, its temperature and density increase and its velocity decreases. As this hot gas then flows around the entry vehicle it can transfer energy to the TPS. This convective heat transfer may be reduced by the use of ablative materials, where the pyrolysis gases evolved from the TPS tend to keep the hot atmosphere gases away from the surface of the material.

The stagnation point convective heat flux may be estimated from the Sutton and Graves [6] correlation $q_{conv} = k \sqrt{\frac{\rho}{r_{nose}}} V_{\infty}^3$ where ρ is freestream density, r_{nose} is the nose radius, and V_{∞} is the flight velocity. The coefficient k ($\text{kg}^{0.5}/\text{m}$) has values as follows: Mars 1.898×10^{-4} , Venus 1.986×10^{-4} , Titan 1.741×10^{-4} , and Earth 1.762×10^{-4} .

At high speeds (several kilometers per second), the strength of the shock wave is such that molecular dissociation takes place. This causes emission of radiation at wavelengths that are characteristic of the atmospheric gases. A proportion of this radiation then reaches the surface of the TPS. The radiative heating of the TPS is thus dependent on the velocity of the entry vehicle and the atmosphere into which it is penetrating. Radiative heat transfer may be computed using the Tauber-Sutton radiative heating correlation for Earth and Mars [7].

The objective of the TPS material is to dissipate as much of this energy as possible and to absorb the remainder.

18.3.3 Thermal Protection System Materials

All thermal protection systems use materials that are designed to minimize the conduction of heat to the entry vehicle. Combine materials combine a low thermal conductivity, high specific heat capacity, and endothermic reactions (pyrolysis) in differing proportions.

Some materials such as the silica tiles used on the Space Shuttle orbiter rely entirely on low conductivity and high heat capacity. Since they do not degrade during entry, they may be used many times. Other materials act as insulators until a specific heat flux is reached, at which point they start to ablate in order to dissipate additional energy. Controlled ablation is a good mechanism for dissipating energy; however, if the ablation is too fast, the heat shield becomes less effective.

A good ablator material will undergo pyrolysis, thereby absorbing energy and releasing gases that transport the energy away from the TPS. Material that has already reacted should remain as a surface char layer in order to act as an additional insulator between the incident heat flux and the pyrolyzing layer. This char layer can be destroyed if the rate of pyrolysis is excessive or the aerodynamic shear on the surface is too great.

Ablators are generally categorized by their density. High-density ablators are capable of withstanding high heat fluxes and producing strong char layers capable of resisting aerodynamic shear forces; lower density ablators cannot survive the most extreme conditions but provide a much lighter design in the conditions for which they are usable.

For the highest heat flux entries, high-density ablators such as carbon-phenolic are required, as in the case of the NASA Galileo probe [1], which had an entry heat flux of approximately 30 kW/cm^2 , and the NASA Pioneer Venus probes. The material used for these probes is no longer in production and no other material with similar properties is currently available. For lower heat flux entries such as Earth sample return missions (Stardust, Genesis), a medium density ablator such as PICA (phenolic-impregnated carbon ablator) may be considered. For the lowest heat fluxes such as Mars entries and entries from orbit, low-density ablators such as SLA561 or Norcoat Liege are more common.

Higher density ablators may often be substituted for lower density alternatives. The design will be less mass efficient but since the ablation will be lower, they may potentially be reusable. For instance, a modified version of PICA is used on the SpaceX Dragon capsule for return from Earth orbit. Since the heat flux in this case is below the pyrolysis threshold for this material, the heat shield is able to be reused.

When designing an entry system, the availability of TPS materials must be considered [8], recognizing that many of the materials used on past missions are no longer manufactured.

18.3.4 Entry Vehicle Geometry

Entry vehicles can be grouped into two general classes: lifting bodies and blunt bodies. Lifting bodies such as the Space Shuttle orbiter, Buran (Russian: Буря́н), and the X-37 use thrusters for attitude control during the early stages of entry and subsequently aerodynamic control. Blunt bodies

such as planetary landers and crewed orbital capsules rely on either thrusters or gyroscopic attitude control during the early stages of entry and aerodynamic stability subsequently.

The need for stability results in stringent requirements on the position of the center of gravity for non-lifting vehicles. The requirements differ with the geometry of the vehicle. Several geometries have been used.

The 70° sphere-cone has been adopted for European and US Mars entry vehicles. It offers the highest drag coefficient of the sphere-cone vehicles (1.6 at Mach 4) but the least static stability. In order for the vehicle to be stable through entry, the center of gravity must be further forward, relative to the overall diameter, than any other sphere-cone.

The 60° sphere-cone was used on the ESA Huygens probe for Titan and the NASA Stardust sample return mission. It has a slightly lower drag coefficient than the 70° sphere-cone (1.5 at Mach 4) but increased stability.

The 45° sphere-cone has a drag coefficient of only 1.1 at Mach 4; however, it is more stable than the previous shapes. This shape was used for the NASA Galileo and Deep Space 2 missions. On the latter, the aeroshell was designed to align itself during entry from any orientation.

The Apollo command modules and ESA's Atmospheric Reentry Demonstrator (ARD) mission used a blunted cone shape which was trimmed to fly at constant angle of attack. Each of these shapes are shown in Fig. 18.3.

18.3.5 Aerodynamics

The aerodynamics of the entry vehicle must be understood in order to predict the trajectory and to ensure its stability. In typical terrestrial applications, the static and dynamic aerodynamic coefficients are functions of the Reynolds number, introduced in Chap. 5, ($Re = \frac{\rho U_\infty L}{\mu}$, where ρ is the density of the flow, U_∞ is the freestream velocity, and μ is the dynamic viscosity) and the Mach number only. As the speed increases to hypersonic, however, the energy of the flow is sufficient to cause molecular dissociation. At this stage, the aerodynamic coefficients are now influenced by the chemical composition of the atmosphere. Thus, an entry vehicle may have different aerodynamic characteristics on Earth and Mars, even at the same Mach number.

As the atmosphere becomes thinner, the flow then transitions from continuum to free-molecular flow as the distance between the individual atmosphere molecules increases. At this stage, the flow then varies with the Knudsen number, introduced in Chaps. 4, 5, ($Kn = \frac{\lambda}{L}$, where λ is the mean free-path between molecules and L is the characteristic length of the vehicle).

The variability of the aerodynamic properties results in the need for a comprehensive aerodynamic database (or databases) covering the chosen vehicle geometry across all

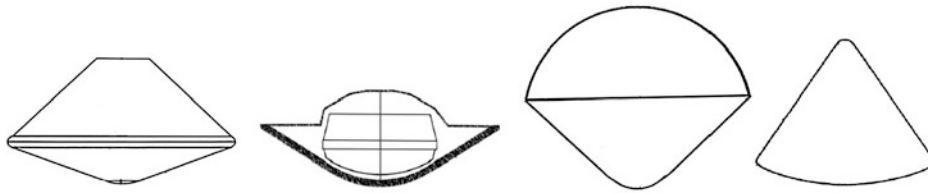


Fig. 18.3 70°, 60° and 45° sphere-cones and Apollo blunted cone

flow regimes in the chosen atmosphere. Such databases, generated by a combination of computational fluid dynamics (CFD) and testing, require significant investment to produce. Thus, only a few shapes have been used for the missions that have flown and aerodynamic databases have been reused.

18.3.6 Inflatable Decelerators

Conventional entry systems use rigid heat shields that must be small enough to fit within the fairing of the launch vehicle. This places a lower limit on the achievable ballistic coefficient and hence on the entry heat flux for a given mission. If the ballistic coefficient could be reduced, this would allow the entry heat flux to be reduced because greater acceleration could take place at high altitude in the lower density atmosphere.

One method of achieving this reduction in ballistic coefficient is to use an inflatable aeroshell. This generally takes the form of a tension cone with its shape maintained by a pressurized torus around its periphery. This technology has been investigated in Europe in the Inflatable Reentry and Descent Technology (IRDT) [9] tests and in the USA with the NASA Hypercone [10]. Although partially successful demonstrator missions have been flown, no mission has yet made use of this technology.

18.3.7 Trajectories

Three general types of trajectory can be flown: ballistic, lifting, and controlled.

The ballistic trajectory is the simplest and most common. The entry vehicle is usually axisymmetric and trimmed to fly at zero angle of attack. During the coasting phase from carrier separation to entry, the entry vehicle is spin-stabilized. The vehicle spin axis is orientated prior to entry such that it will be aligned with the air-relative velocity vector at the time that sufficient aerodynamic force is generated to facilitate control of the vehicle attitude during entry. The spin continues throughout the entry, in order to cancel out the effects of any asymmetry in the outer geometry of the vehicle on the aerodynamics and any error in center of

gravity position on the motion. This entry technique needs no guidance system, and is not controllable. The descent system is generally triggered by an accelerometer that estimates the instantaneous velocity from its decelerating effect on the vehicle.

A ballistic trajectory is suitable for planetary entry missions where simplicity and robustness are more important than accurate targeting.

The lifting trajectory is a variation of the ballistic trajectory. The entry vehicle is usually an axisymmetric shape with an offset center of gravity (such as Apollo or ARD). Since the vehicle will naturally fly at a constant angle of attack, it will generate lift as well as drag. For this type of trajectory, the lift to drag ratio is constant; the trajectory is always lifting and so if no lift is required, frequent roll reversal maneuvers are necessary.

In order to make use of this lift, the entry vehicle must incorporate a control system and a means of reorientating the lift vector. The control system generally uses accelerometers to measure the deceleration of the vehicle and an inertial platform to determine its orientation. The lift vector can be reoriented by using thrusters to rotate the vehicle about its axis.

A lifting vehicle allows a control system to tailor the trajectory in several ways. The lift vector can be rotated upwards or downwards. This allows the trajectory to move between more or less dense regions of atmosphere in order to maintain either a constant density or a constant acceleration. For a crewed mission, the deceleration can be limited to a human tolerance level by ascending into the more rarefied atmosphere if the deceleration becomes too large. Alternatively, a shallow trajectory may be chosen in order to limit the heat flux but the lift vector may be used to preclude skipping out of the atmosphere.

More complex guidance algorithms such as that used by Apollo allow the roll position to be used to control the parachute deployment position. Cross-track adjustments may be made by biasing the roll position to one side or the other, and downrange adjustments may be made by ascending into the more rarefied atmosphere to increase the downrange distance or by descending into the denser atmosphere to reduce it.

As with ballistic entry vehicles, the lifting vehicle is likely to be spin stabilized during the coast phase. Just

before entry, the spin must be neutralized using a de-spin system. The guidance system then controls the roll position of the spacecraft in order to follow the desired trajectory.

Controlled trajectories such as that of the Space Shuttle allow the lift vector to be rotated, as with the lifting trajectory, and also its magnitude to be varied. This allows greater control over the entry trajectory at the cost of a more complicated control system. In this case, the vehicle attitude must be controlled using thrusters during the initial phase of entry until sufficient aerodynamic force has built up to allow the control surfaces to perform their functions.

18.3.8 Transition to Descent

The vehicle shapes used for ballistic or lifting entries are dynamically unstable at transonic velocities. Consequently they could not be used for an uncontrolled descent to the surface, even if the velocity was acceptable. The entry phase thus usually ends at a Mach number that is no less than about 1.4.

Since it is impossible to measure Mach number directly, it must be inferred using alternative means. The most common method of determining the correct time to start the descent sequence is by using an accelerometer to detect the deceleration of the vehicle. If the atmospheric structure and drag of the entry vehicle are well known, the trajectory may be predicted and the Mach number correlated with acceleration. An alternative approach is to use an inertial platform to monitor the deceleration through descent; however, the initial setting of the platform must be very accurate and the winds at the deployment location must be low in order to avoid errors.

18.4 Descent

The descent system controls the spacecraft from the end of entry through to contact with the surface. The purpose of the descent system is to reduce the velocity relative to the planet's surface to a value compatible with the landing system, to stabilize the vehicle, and possibly also to guide the vehicle to its landing point. Generally, the descent system comprises parachutes and/or a propulsive system. Gliding vehicles (Space Shuttle orbiter), auto-rotating helicopters, and balloons may also be considered as descent systems.

18.4.1 Heritage Descent Systems

Descent systems that have been used on robotic Western space missions are summarized in Table 18.1 using data from "Aerodynamic Decelerators for Planetary

Exploration: Past, Present and Future" [11] supplemented with some limited data on Russian missions. Similar data for crewed missions is shown in Table 18.2 [12, 13].

All successful Mars landers have used a single parachute followed by either a retro-rocket or a propulsive descent system. Such a system is selected because it would require a very large and grossly mass-inefficient parachute system to land a probe without propulsion. Added to this, there would be severe risk of the parachute enveloping the lander. A two-stage parachute system can provide a lower mass solution but, although studied for both ExoMars 2016 and Mars Science Laboratory (MSL), this sequence has been deemed too complex.

Both Pioneer Venus and Galileo [14] used staged systems in order to aid separation of the aeroshell from the probe. Huygens selected a similar approach, but added a third parachute to the sequence to allow the lander to reach the surface of Titan in the required time.

All Earth return systems also use staged parachute sequences, with a drogue parachute deployed to stabilize the vehicle and reduce the dynamic pressure prior to deployment of the terminal descent parachute.

Western crewed systems have all employed water landings and so no propulsive system is used to reduce the impact velocity. The Russian Soyuz descent system, which returns to land, uses a retro-rocket to reduce the impact velocity.

18.4.2 Parachutes

Parachutes play a key role in the entry, descent, and landing of space vehicles. Parachutes provide deceleration, often from supersonic to subsonic speed. They can also achieve a specific descent rate to enable scientific measurements to be obtained, provide stability (drogue function) to prevent the aeroshell from tumbling, to meet instrumentation requirements, to deploy another parachute (pilot function), to achieve the ballistic coefficients required to enable separation events, and to pursue a trajectory suitable for completion of the EDL sequence.

There are many types of parachute but they fall into two functional groups: ballistic parachutes designed to provide drag and stabilization, and steerable gliding parachutes designed to provide some degree of control over the landing site during descent. To date all extra-terrestrial probes and crewed missions have used ballistic parachutes (Fig. 18.4). Gliding parachutes have been used for one Earth return mission: Genesis [15]. A gliding parachute was also developed for the planned X-38 crewed system [16] (Fig. 18.5).

Table 18.1 Descent and landing systems for robotic extra-terrestrial and Earth return missions; numbers in parenthesis indicate stages

Mission	Parachute type	Diameter or area	Deployment Mach	Deployment system	Terminal descent	Landing system
Viking	Disk-gap-band	16.2 m D_0	2.1	Mortar	Propulsive	Landing legs
Pioneer Venus	(1) Ribless guide surface	(1) 0.76	(1) 0.8	(1) Mortar	None	None
	(2) Conical ribbon	(2) 4.94 m D_0	(2) 0.8	(2) Pilot parachute		
Galileo	(1) Conical ribbon	(1) 1.14 m D_0	(1) 0.95	(1) Mortar	None	None
	(2) Conical ribbon	(2) 3.8 m D_0	(2) 0.95	(2) Pilot parachute		
Mars 2, 3, 6, 7	(1) Cross	(1) 4 m D_0	(1) 3.5	(1) Mortar	Retro-rocket	None
	(2) Reefed cross	(2) 13.4 m D_0	(2) 3.4	(2) Pilot parachute		
Mars Pathfinder	Disk-gap-band	12.7 m D_0	1.71	Mortar	Retro-rocket	Non-vented airbags
Mars Polar Lander	Disk-gap-band	12.7 m D_0	1.7–1.85	Mortar	Propulsive	Landing legs
Beagle 2	(1) Disk-gap-band	(1) 3.2 m D_0	(1) 1.5	(1) Mortar	None	Non-vented airbags
	(2) Ringsail	(2) 10.0 m D_0	(2) 0.4–0.6	(2) Pilot Parachute		
Mars Exploration Rovers	Disk-gap-band	14.1 m D_0	1.8 and 1.9	Mortar	Retro-rocket + Transverse Impulse Rocket System, TIRS	Non-vented airbags
Huygens	(1) Disk-gap-band	(1) 2.6 m D_0	(1) 1.47	(1) Mortar	None	None
	(2) Disk-gap-band	(2) 8.3 m D_0	(2) 1.36	(2) Pilot parachute		
	(3) Disk-gap-band	(3) 3.0 m D_0	(3) 0.15	(3) Pilot parachute		
Genesis	(1) Disk-gap-band	(1) 2.03 m D_0	(1) 1.8	(1) Mortar	Mid-air retrieval	None
	(2) Parafoil	(2) 325 m ²	(2) subsonic	(2) Pilot parachute		
Stardust	(1) Disk-gap-band	(1) 0.8 m D_0	(1) 1.4	(1) Mortar	None	None
	(2) Triconical	(2) 7.3 m D_0	(2) 0.15	(2) Pilot parachute		
Phoenix	Disk-gap-band	11.7 m D_0	1.3	Mortar	Propulsive	Landing legs
Mars Science Laboratory, MSL	Disk-gap-band	21.5 m D_0	2.0	Mortar	Propulsive	Sky-crane
ExoMars 2016	Disk-gap-band	12.0 m D_0	1.8–2.1	Mortar	Propulsive	Crushable

18.4.2.1 Basic Parachute Physics

The most significant parameter in the design of a ballistic parachute is its drag coefficient, i.e. the component of aerodynamic force resolved in the direction of the air flow. The deceleration of the parachute and payload is achieved by transferring its momentum to the air through which it passes. The drag, D , is related to the drag coefficient, C_{D_0} , the canopy area, S_0 , and the dynamic pressure, $q = (\rho V^2/2)$, by Eq. 4.105, which can also be written as

$$D = \frac{1}{2} \rho V^2 C_{D_0} S_0 \quad (18.1)$$

The product $C_{D_0} S_0$ is referred to as the drag area, and is used to define parachute performance requirements for a given application. The canopy area, S_0 , is the entire surface area of the canopy including any openings or vents. Typically, C_{D_0} has a value between 0.4 and 0.8 and is influenced by several parameters: porosity, canopy shape, suspension line length, payload wake effects, the Reynolds number per unit length and the Mach number. Of these, porosity is the most important.

Two forms of porosity are introduced into a parachute canopy. Geometric porosity is produced by physical gaps in the canopy and permeability from the air flow passing

Table 18.2 Descent and landing systems for crewed missions; numbers in parenthesis indicate stages

Mission	Parachute type	Diameter or area	Deployment Mach	Deployment system	Terminal descent	Landing system
Mercury	(1) Disk-gap-band	(1) 2.6 m D_0	(1) 1.47	(1) Mortar	None—water landing	Vented airbag
	(2) Disk-gap-band	(2) 8.3 m D_0	(2) 1.36	(2) Pilot parachute		
	(3) Ringsail	(3) 3.0 m D_0	(3) 0.15	(3) Pilot parachute		
Gemini	(1) Reefed conical ribbon	(1) 2.6 m D_0	(1) 0.84	(1) Mortar	None—water landing	None
	(2) Reefed ringsail	(2) 5.5 m D_0	(2) 0.27	(2) Pilot parachute		
	(3) Reefed ringsail	(3) 25.7 m D_0	(3) 0.26	(3) Pilot parachute		
Apollo	(1) Ringslot	(1) 2.2 m D_0	(1) 0.70	(1) Mortar	None—water landing	None
	(2) Reefed conical ribbon ($\times 2$)	(2) 5.9 m D_0	(2) 0.70	(2) Mortar		
	(3) Ringslot ($\times 3$)	(3) 2.2 m D_0	(3) 0.23	(3) Mortar		
	(4) Reefed ringsail ($\times 3$)	(4) 26.0 m D_0	(4) 0.23	(4) Pilot parachute		



Fig. 18.4 A parachute for the Galileo spacecraft is tested in a wind tunnel at NASA Langley Research Center. Galileo used a conical ribbon, ballistic parachute. *Image* NASA; GPN-2000-001911 (18 April 1983)

through the canopy fabric. Geometric porosity is the ratio of the open areas of the canopy to the total area of the canopy. The performance of a parachute is affected by the total porosity of the canopy, both geometric and fabric. Increasing the porosity reduces the drag coefficient, reduces the trim angle, increases the static stability, increases the

parachute inflation time, and reduces the parachute inflation loads. The effect of porosity on the drag coefficient (adapted from [17]) is shown in Fig. 18.6, while Fig. 18.7 shows its effect on the trim angle.

A parachute with a trim angle of 0° angle of attack will always descend vertically, however a parachute with a trim angle other than zero will descend in one of three modes: oscillation, coning, or gliding. The mode depends on a parameter called the mass ratio, which is the ratio of the mass of fluid associated or added mass with the canopy or to the payload mass. This may be written $(2.136 \rho (\pi D_p^3 / 12) / m_s)$, where m_s is the payload mass and D_p is the canopy projected diameter. With a high mass ratio the payload mass is small compared to the mass of fluid associated with the canopy, and the center of gravity of the system is near the canopy. In these circumstances, the system glides at the trim angle to the vertical. When the mass ratio is low, the system center of gravity is close to the payload and the system either oscillates or cones. This effect can be very important in the case of planets with low-density atmospheres. A 10 m diameter parachute may have a mass of only a few kilograms; however, the added mass will be as much as 200 kg on Earth. Thus, the parachute motion is dominated by the added mass rather than the physical mass of the parachute. On Mars, the situation is different with the added mass for the same parachute being only about 1.5 kg, which is less than the mass of the parachute itself. A parachute that glides on Earth is likely to be unstable in a low-density atmosphere. This was demonstrated clearly during the testing of a low porosity ringsail parachute in the early development tests for Mars Science Laboratory (MSL) [18]. A high lift-to-drag gliding

Fig. 18.5 The X-38 Crew Return Vehicle lands at the end of a July 1999 test flight at the Dryden Flight Research Center using a steerable gliding parafoil. *Image* NASA; NIX-EC99-45080-101 (1 July 1999)



Fig. 18.6 Effect of porosity on parachute drag coefficient

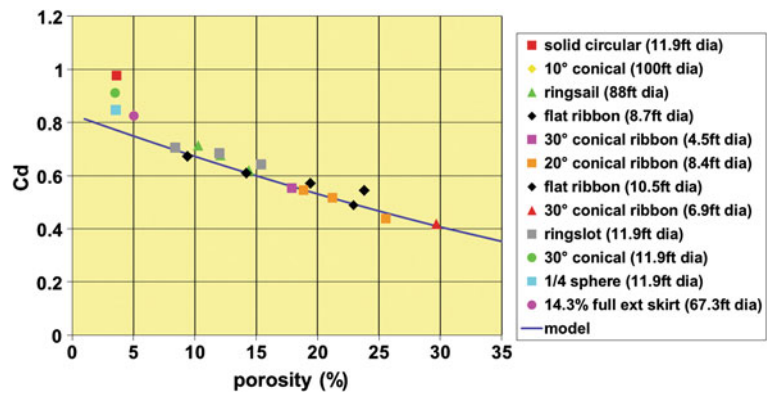


Fig. 18.7 Effect of porosity on trim angle

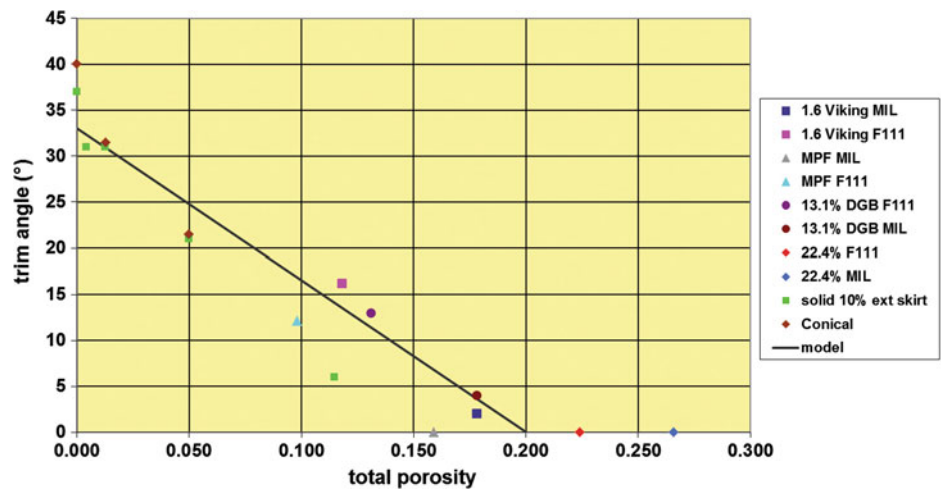
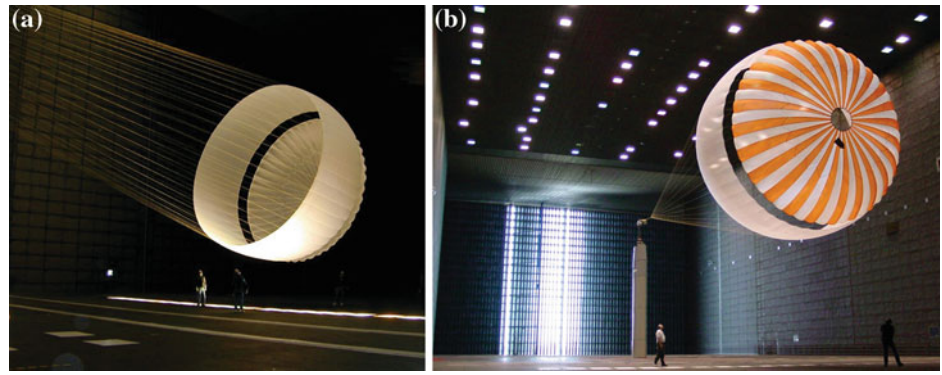


Fig. 18.8 Extended skirt on the Mars Exploration Rover (MER) parachute in wind tunnel at NASA Ames Research Center



parachute is unsuitable for Mars operation, since any control input will simply cause an overly rapid response of the parachute above the payload.

Canopy shape only has a small effect on the drag coefficient except for one feature that is called an extended skirt, shown in Fig. 18.8. The extension is either a cylinder or an inverted frustum whose angle matches that of the suspension lines and it acts to enhance stability at the expense of the drag coefficient because the extension does not add to the drag-producing surface. This feature was employed by Mars Pathfinder [19] and Mars Exploration Rover [20] to provide the high stability necessary for the retro-rockets.

The Reynold's number per unit length affects the permeability of porous broadloom material. At low values ($<5 \times 10^4$) the permeability is very low. At higher values ($>5 \times 10^6$) the permeability tends to its maximum value for the specific material. For low density atmospheres, like Mars, (R_e/m) tends to low values whereas at low altitude on Earth it is quite high. Parachutes constructed from porous broadloom could therefore change performance [21]. Based on this, it is recommended to use zero porosity broadloom in order to maintain consistent behavior.

Line length also has an effect on the parachute drag coefficient. Typically, parachute line length is 0.8–1.0 D_0 . Shorter lines can reduce the drag coefficient. Longer lines increase the drag coefficient for parachutes whose shape is largely determined by the pressure distribution (flat circular) but they have only a small influence on parachutes with a shape largely determined by the design, such as poly-conical designs.

As the parachute always flies in the wake of its payload, this creates an inextricable link between the two, and parachute drag is diminished by the reduced fluid momentum in the wake. For large parachutes and small payloads, the reduction in drag of the parachute due to this mechanism is approximately equal to the drag of the payload. Moreover, the unsteady wake of the payload can cause the parachute to move in sympathy to the changing flow, stimulating oscillations.

The Mach number also has a significant effect on the drag coefficient. Parachutes that perform well in the

subsonic range, such as conical ribbon parachutes, suffer a rapid reduction in drag performance at quite low supersonic speeds. The subsonic drag coefficient is approximately constant below Mach 0.4, subsequently rising steadily to Mach 0.75 in the normal way for a bluff body in a compressible flow. Between Mach 0.75 and Mach 1.0 the drag coefficient may reduce. From Mach 1.0 to Mach 1.4 the drag coefficient remains reasonably constant before falling steadily above Mach 1.5. Two mechanisms may cause cyclical inflation and collapse of the parachutes in a supersonic flow. One derives from flying the parachute too close to the base of the vehicle. This can couple the subsonic region of the wake with the subsonic region in the parachute canopy, with consequential pressure bleed. Consequently, supersonic parachutes should be designed to fly well downstream of the payload. A trailing distance $L_T > 6D_B + D_0$ where D_B is the payload diameter has been proposed [22]. Another mechanism occurs even for a parachute that is deployed far enough behind the probe to avoid pressure coupling. It results from the disturbances in the payload wake that cause parachute bow shock movements, pressure fluctuations in the parachute, and gross shape changes. The payload wake stimulates the bow shock to move cyclically close to the parachute mouth (high pressure and full inflation) and well upstream of the parachute mouth (low pressure and partial collapse). The onset of this phenomenon is between Mach 2 and Mach 2.4. This behavior has resulted in an upper limit of Mach 2.3 being set for the deployment of the disk-gap-band parachute used for all Western missions to Mars.

Rather than using a single parachute, several identical parachutes can be deployed in a cluster, as shown in Figs. 18.9 and 18.10. Each parachute will contribute to the overall drag of the system. Clustering allows the use of several smaller parachutes that may be easier to deploy, rather than one large one. This also provides some degree of redundancy in the case of one canopy failing—this was the principle for Apollo [13], and is shown in Fig. 18.10. The disadvantages include system complexity and the greater inflation loads associated with faster opening.

Fig. 18.9 An Ares I main cluster parachute test at the U.S. Army Proving Grounds in Yuma, Arizona; May 20, 2009. It involved cluster of three 45 m (150 feet) diameter parachutes lowering a test weight to the desert floor. *Image* U.S. Army Proving Ground



When flying in a cluster, the parachutes do not touch each other but naturally leave a gap between them (Fig. 18.9). Parachutes flying in a cluster usually provide lower drag than they would individually. However, parachutes with large trim angles fly stably in a cluster and hence it is possible to take advantage of their high drag coefficients.

The highest load applied to a parachute occurs when it inflates, and this defines the strength needed for the parachute materials. The inflation process can be broken down into a number of steps. Inflation begins when the parachute mouth opens to accept a single ‘breath’ of air. This breath will then pressurize the top of the crown of the parachute. This pressurization will cause the mouth of the parachute to open, thereby allowing more air to flow into the parachute. This air will cause the inflated portion of the parachute to increase from the crown towards the hem, until the canopy is fully inflated. The residual radial momentum in the inflating parachute may cause over-inflation to occur before the parachute settles into the steady-state inflated condition. For a parachute that inflates at close to constant velocity, for example in a wind tunnel test or when the mass ratio is small, the maximum load occurs close to the end of inflation, when the parachute diameter is increasing less rapidly and the rate of change of momentum of the gas around the parachute is greatest. In this case, the ratio of the peak drag-coefficient during the inflation to the steady state drag coefficient is called the inflation factor. This factor typically is 1.1–2.0 depending on the parachute design and its porosity. When the parachute mass ratio is large and the parachute decelerates rapidly during inflation, the maximum load will occur earlier in the inflation.

Prediction of parachute inflation is a particularly complex problem. The inflation process is influenced by many factors including flow velocity, atmosphere density, Mach number, payload wake, material, payload and canopy masses, and the manner in which the parachute is packed. References [23–27] present various approaches.

Reefing is a technique used to modulate the drag of a parachute during opening and initial deceleration in order to control both the force imposed on the payload and the stresses within the parachute. The most common technique is to sew a small metal ring onto each parachute line where it crosses the canopy mouth, and then pass a reefing line through all the rings. The line will have a shorter length than the fully open circumference of the parachute and will thus constrict the parachute. Pyrotechnic cutters are provided to release the line after sufficient time has passed to allow the payload to decelerate and the parachute force to reduce.

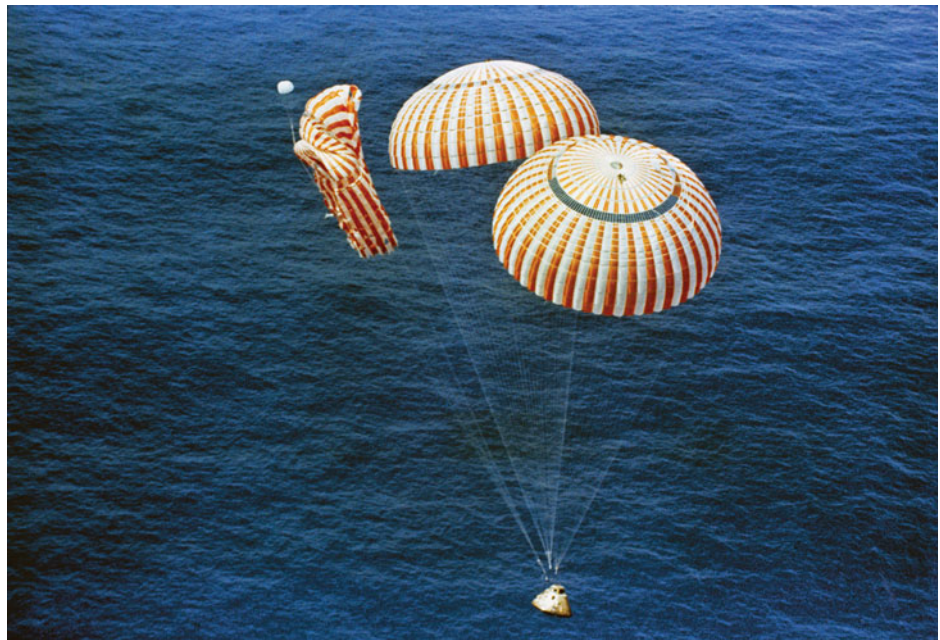
18.4.2.2 Disk-Gap-Band

Various configurations of the disk-gap-band (DGB) parachute have been used as the supersonic or single parachute on all US Mars landers and the European Huygens and Beagle-2 probes. This parachute was selected by NASA for the Viking probe owing to its inflation and flight characteristics at high Mach number in low density atmospheres [28]. It was also used as both the supersonic and the subsonic parachute on the ESA Huygens mission [29] (Fig. 18.11).

Structurally, the parachute consists of a flat circular disk, a cylindrical gap, and a cylindrical band.

Several designs of DGB have been used, and each has different characteristics. The Viking mission used a

Fig. 18.10 The capsule of Apollo 15 descending safely despite a parachute line failure. *Image* NASA; AP15-S71-42217 (7 August 1971)



parachute with a geometric porosity of 12.5 % distributed between the vent (0.5 %) and gap (12 %). For the Mars Pathfinder (MPF) mission, much greater stability was required than the 15-degree trim angle of the Viking design in order to ensure optimum performance of the retro-rockets. Its stability was improved by increasing the length of the band portion by a factor of 1.9 over that of the Viking design [19]. The ESA Huygens mission also required very high stability for the on-board camera. The solution adopted was to increase the width of the gap by a factor of 2, increasing the geometric porosity to 22.4 %, and providing a design with a trim angle close to zero degrees [29]. Stardust, Mars Polar Lander, Genesis, Phoenix, and MSL all used the Viking geometry. The Mars Exploration Rover (MER) program used a very similar parachute to that of MPF, with a band length 1.8 times the Viking value [20].

The low subsonic drag coefficients for these disk-gap-band parachutes were typically 0.41 for MPF, 0.43 for MER, 0.49 for Huygens, and 0.6 for Viking (Fig. 18.12).

18.4.2.3 Conical Ribbon

Conical ribbon parachutes are manufactured to have a shallow cone shape with the gores constructed from spaced ribbons or tapes, usually 50 mm in width. These geometric porosity is typically 15–30 % with a drag coefficient of 0.6–0.4. They are used extensively on Earth for high dynamic pressures in the transonic and subsonic regimes because their construction method allows very high strength. Ribbon parachutes were used on the Galileo and Pioneer Venus missions where opening at high dynamic pressure (6 kPa) was required. Figure 18.4 shows the wind tunnel model of the Galileo ribbon parachute.

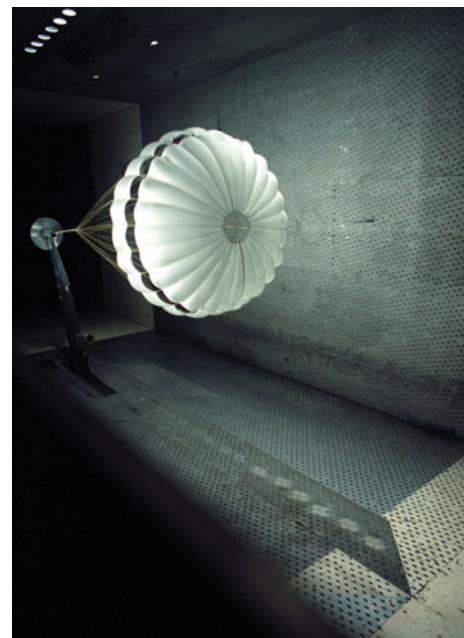


Fig. 18.11 The Huygens disk-gap-band parachute under test in the 16-foot transonic wind tunnel at the US Air Force's Arnold Engineering Development Center in 1993. The wind tunnel model of the probe was fitted with scaled main and pilot parachutes. The chutes were opened at speeds ranging from 350–1,000 mph in the wind tunnel while information was gathered on their inflation characteristics. The Huygens test model was designed and fabricated by Micro Craft Inc. in Tullahoma, Tennessee, for the GE Aerospace Corp., which was under contract to the European Space Agency. *Image* Arnold Air Force Base

18.4.2.4 Ringslot

A ringslot parachute is a conical canopy constructed from bands of broadloom fabric with gaps, and is rather like a

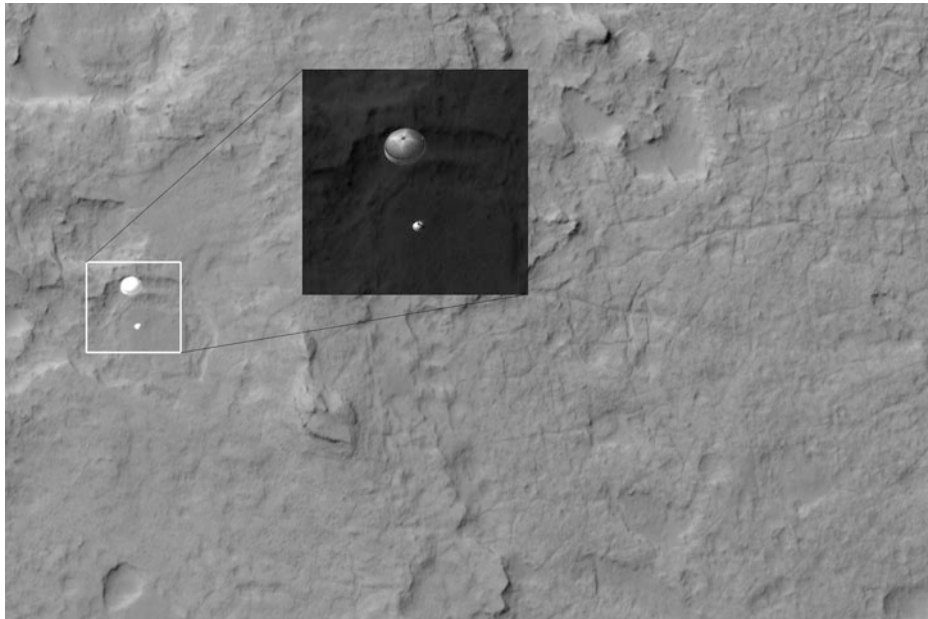


Fig. 18.12 The Mars Science Laboratory (MSL) Curiosity rover and its parachute viewed from the Mars Reconnaissance Orbiter's High-Resolution Imaging Science Experiment (HiRISE) as Curiosity descended to the surface on Aug. 6 (UTC). Curiosity and its parachute are in the center of the white box; and a separate image is a smaller cutout of MSL stretched to avoid saturation. The rover is landing on the etched plains just north of the sand dunes that fringe Mt. Sharp.

The parachute appears fully inflated and performing well. Details in the parachute such as the band gap at the edges and the central hole are clearly visible. The cords connecting the parachute to the backshell cannot be seen, although they were seen in the image of Phoenix descending, perhaps due to the difference in lighting angles. *Image* NASA/JPL/University of Arizona/HiRISE Team

course ribbon parachute. Its porosity is typically in the range 15–20 %, its drag coefficient is in the range 0.50–0.57, and its trim angle is close to zero degrees at the lower drag coefficients. It has a good heritage in the subsonic regime and was used in the Apollo system for pilot chutes. It was also studied within the ESA ExoMars mission as a candidate design for subsonic terminal descent (Fig. 18.13).

18.4.2.5 Ringsail

The ringsail parachute is essentially a ringslot design with increased fullness in the lower edge of the bands away from the vent region. It was used for the Apollo, Gemini, and Mercury main parachutes. Ringsail parachutes were tested in the NASA Planetary Entry Parachute Program (PEPP): a series of high altitude, supersonic, full scale tests which were a precursor to the Viking mission [28]. The parachutes tested had a porosity of 16–17 %, and drag coefficients of 0.6 were measured with trim angle less than plus or minus 10°.

This parachute appears to have a very high drag coefficient of up to 1.2 at low altitudes on Earth [18], owing to parachute glide. At higher altitudes on Earth, or at any altitude on Mars, the parachute is unstable and a much lower drag coefficient is observed. A low porosity ringsail parachute was baselined as the main parachute for MSL but was later discarded following when it proved to be unstable in flight at high altitude.

18.4.2.6 Cruciform

Cruciform or cross parachutes were tested as part of the preparation for the Viking mission [28] and proved to be unsuitable for supersonic operation. The drag coefficients for these parachutes found in literature are higher than those for axisymmetric parachutes; however, this is due to the use of the fabric area rather than the nominal area to calculate the drag coefficient. In subsonic flow, they exhibit similar drag performance, based on the cloth area, to disk-gap band parachutes but they have a rotational instability. The stability of cross parachutes varies with the ratio of arm length to width (i.e. the aspect ratio). Both the drag coefficient and the trim angle reduce with increasing aspect ratio.

In Russia, cruciform parachutes have been utilized where the extremities of the cross shape are connected by a band of material. The addition of this material prevents the supersonic scissoring motion exhibited by the standard cruciform parachute, but makes the parachute less drag efficient since additional canopy surface is required without any increase in drag.

This parachute has been used within the Russian space program; however, little information is available in the open literature. A version of the modified cruciform was studied in the early phases of ExoMars and was shown to have a low drag coefficient similar to the Mars Pathfinder DGB of 0.41.

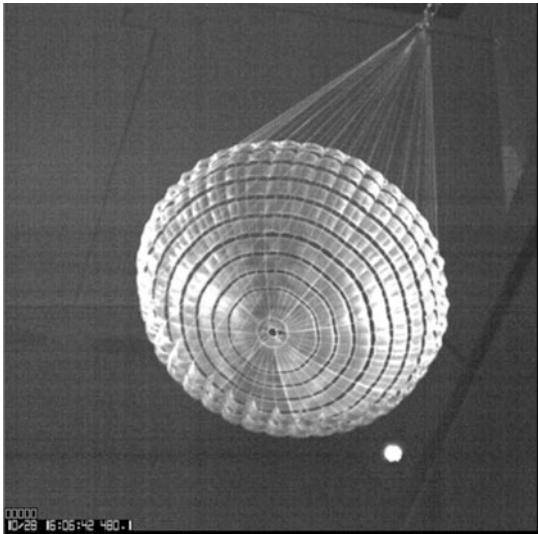


Fig. 18.13 Video screen-grab of ExoMars ringslot parachute during wind tunnel testing

18.4.2.7 Polyconical

Polyconical parachutes are constructed using broadloom fabric gores. The gores are patterned to give a constructed geometry consisting of a number of concentric conical sections. They are usually constructed using very low porosity material.

Polyconical parachutes appear to have drag coefficients of 1.2 or more at low level on Earth, based on their descent velocity. In reality, from Fig. 18.7, the total velocity V is greater than the descent velocity, with $V = (V_V / \sin 2\alpha)$ where V_V is the descent velocity and α is the trim angle; the drag coefficient is really about 0.8 and the trim angle is 30° to 40° to the vertical.

Polyconical parachutes are frequently modified by the addition of drive slots, symmetrical openings in the gores located to the rear of the canopy, in order to provide directional glide. In practice, these slots do not cause the parachute to glide; they simply define the plane in which it glides naturally. Typical applications are the descent parachute for the Stardust probe [30] and ejection seat parachutes.

18.4.2.8 Parafoil

The parafoil [31], when inflated, resembles a low aspect ratio wing with a typical aspect ratio of 3.0. It is entirely constructed from fabric with no rigid members, which allows it to be packed and deployed in a manner similar to a conventional parachute canopy. The wing has upper and lower membrane surfaces, an airfoil cross section, and a rectangular planform. The airfoil section is formed by airfoil shaped ribs sewn chord-wise between the upper and lower membrane surfaces at a number of span-wise

intervals to form a series of cells. The leading edge of the wing is open over its entire length so that ram air pressure will maintain the wing shape. The fabric used in the manufacture of ram-air parachutes is as imporous as possible in order to obviate pressure loss. The suspension lines are generally attached to alternate ribs at multiple chord-wise positions. Although this results in a large number of suspension lines, it is necessary in order to maintain the chord-wise profile. Parafoils can glide with a lift to drag ratio (L/D) in excess of 3:1. They are maneuverable and can land with pin-point accuracy using on-board guidance and control systems. Parafoils have proved effective from small 2 m^2 wing area devices suitable for munitions right up to the 697 m^2 X-38 parafoil [16] shown in Fig. 18.14.

Parafoils were studied for pinpoint landings on Mars but the results suggested that their response to control inputs would be too violent because of the low added mass on Mars; however, for denser atmospheres they have great potential.

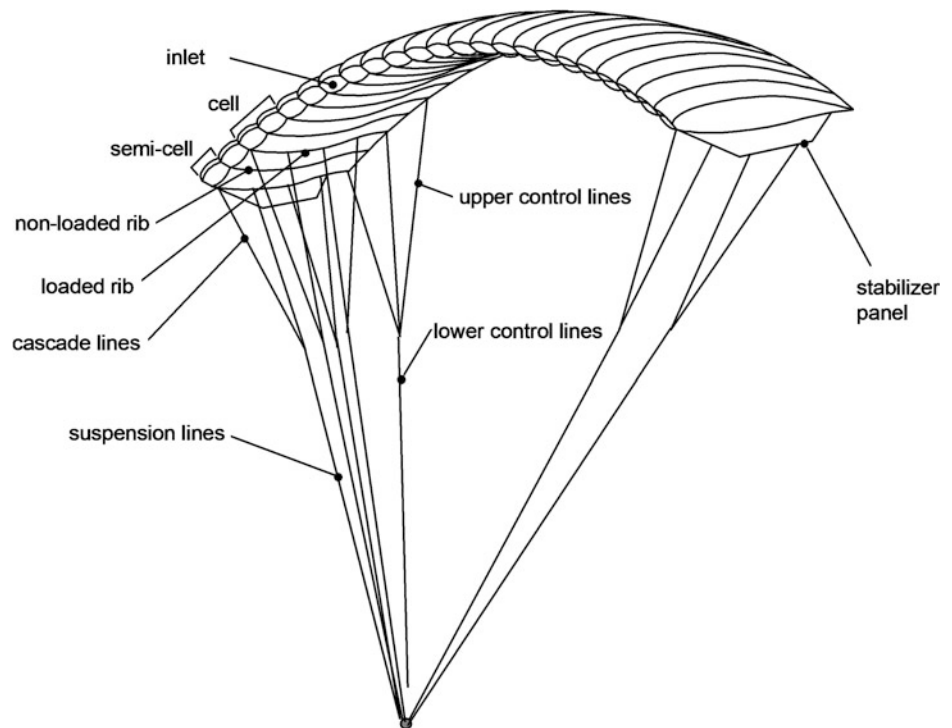
18.4.2.9 Deployment

The first parachute in a descent sequence must be positively deployed in order to pass through the payload wake. Various methods of deployment are possible, including pyrotechnic mortars and tractor rockets.

Pyrotechnic mortars [32] are the usual way of deploying pilot chutes for space missions. The parachute is packed in a tube, open at one end, above a moving piston. A pyrotechnic charge pressurizes the closed volume under the piston and forces the piston out of the tube, pushing the parachute ahead of it. This is a very efficient method of deploying a parachute; however, since the acceleration takes place over a very short distance and thus time, the reaction loads are large for all but small parachutes. For example, a 15 kg parachute ejected at 30 m/s would have a reaction load of about 40 kN.

Tractor rockets are used to deploy larger parachutes. They are used on some ejection seats and capsules, and on whole-aircraft recovery systems. A tractor rocket pulls the parachute out of a container and positively accelerates it through deployment. Since the force acts on the parachute for much longer, the force and acceleration can be much lower. Furthermore, there is no reaction load on the payload. Tractor rockets are not currently used on space missions due to their increased integration complexity.

Second and subsequent parachutes are generally deployed by the released parachute as it separates from the payload. This can present challenges with large parachutes at high velocities, since the released parachute can accelerate the deploying parachute to a high enough velocity during deployment to risk significant deployment snatch loads and friction damage.

Fig. 18.14 Parafoil

18.4.3 Propulsive Descent

18.4.3.1 Descent Propulsion Systems

A descent propulsion system was first used on a crewed vehicle for the Apollo Lunar Module, however it had previously been used on robotic missions including the Surveyor lunar landers; the first vehicles to unintentionally (Surveyor-3) and intentionally (Surveyor-6) liftoff from the lunar surface. The Apollo Lunar Module propulsion system transferred the vehicle, containing two crew, from the circular 110 km lunar parking orbit to a 15 km perilune descent orbit. Subsequently a powered descent to the lunar surface, including sufficient hover time at low altitude to select the exact landing site, was provided. The hypergolic bipropellant propulsion system, which used Aerozine 50 and nitrogen tetroxide, and was throttleable from 10 to 60 % thrust with a maximum 100 % thrust of 45 kN. The Apollo Lunar Module's descent propulsion system (DPS) engine was able to be gimballed to precisely align the thrust vector through the center of gravity at start-up, after which it was held in position and the vehicle changed orientation to steer.

The Viking missions in 1976 were influenced by the design of Apollo. The Viking descent propulsion system [33] brought the lander to within $2.4(\pm 1)$ m/s vertical velocity and <1 m/s horizontal velocity at touchdown. Trajectory and attitude control was achieved through the use of an inertial reference unit, four gyros, a radar altimeter, and a Doppler radar to detect horizontal velocity. The system was initiated at about 65 m/s (200 ft/s) at an altitude

of 1,200 m (4,000 ft.). The three engines orientated the spacecraft so that their combined thrust vector opposed the velocity vector of the spacecraft (i.e. a gravity-turn maneuver). The landing trajectory was maintained between two preprogrammed limiting altitude/velocity profiles. The lander reached a height of about 16.8 m above the surface with a residual velocity of 2.4 m/s and continued to the surface at this velocity. As soon as a sensor on any one of the three footpads touched the surface, the rocket engines were switched off. The system utilized three monopropellant hydrazine engines separated by 120° . The engines had 18 nozzles to diffuse the exhaust plume, eliminate contamination of the spacecraft instrumentation due to recirculation, and avoid erosion of the landing site. The engines were throttleable from 276 to 2,667 N. The 85 kg of propellant was contained in two spherical titanium tanks mounted on opposite sides of the lander.

The Phoenix system [33] was almost identical to the Mars Polar Lander system that ended in loss during its landing attempt in 1999. The system was similar to Viking but included several cost reduction features. A canted multi-beam radar was employed to avoid horizontal Doppler radar velocity measurement. Throttled engines were replaced by twelve, 302 N monopropellant hydrazine thrusters off-pulsed engines operating at high duty cycles.

The Mars Science Laboratory (MSL) landing system [34] took a radical approach to landing system design. For landers of the Viking type, it is important to avoid risks to the system from plume impingement on the regolith. This is achieved by descending as rapidly as the landing legs will

allow. This adversely affects both ground clearance under the vehicle and slope tolerance. Placing the propulsion system under a rover also presents egress issues. The upgrading of MPF/MER propulsion configuration to place throttled monopropellant engines in the backshell above the lander solved the problem. In this way, the descent engines could lower the rover to the surface at a much slower velocity. This system, called Sky-crane, removed the need for heavy landing gear (like airbags) and increased the tolerance of the lander to slopes and rocks. The powered descent was initiated at an altitude between 1,500 and 2,000 m above ground level and at a velocity near 100 m/s. The descent stage comprised eight variable thrust monopropellant hydrazine rocket thrusters in pairs on four arms extending around this platform. Each rocket thruster produced 400–3,100 N of thrust. It was derived from those used on the Viking landers. The MSL lander flew a Viking-like gravity-turn maneuver to reduce the vertical velocity to 0.75 m/s and the lateral velocity to zero at 20 m above the ground. The rover was then lowered on a 7.5 m bridle while the descent velocity was maintained until either touchdown or bridle off-load was detected. The rover was then released and the Sky-crane performed a flyaway maneuver.

18.4.3.2 Retro Rockets

The Mars Pathfinder mission [3] sought considerable cost reductions compared to Viking. Its strategy to reduce cost was to use the Viking entry and descent systems (with passive attitude control) and use low cost solid rocket engines that would deliver the lander to the surface with much larger range of touchdown velocities than legged landers could typically handle, necessitating the development of the tetrahedral airbag (Sect. 18.5.2). This also removed the requirement for horizontal velocity estimation with a Doppler radar. At 1.6 km above the surface, while descending on the parachute, a radar altimeter was used to determine both the altitude and descent velocity; this information was then used by the on-board computer to determine the precise timing of the landing events. The airbags were inflated at 355 m above the ground. The three solid retro-rockets mounted above the lander in the backshell were fired at a height of 98 m. When the computer estimated the lander was stationary at a height between 15 and 25 m the bridle was cut and the payload fell to the ground, protected by the airbag system. The rockets flew up and away with the backshell and parachute.

The Mars Exploration Rovers [35] employed a similar retro-rocket system, but enhanced with a simplified form of horizontal velocity control in the form of the transverse impulse rocket system (TIRS). The TIRS does not provide a means of countering steady-state wind velocity, it is designed to mitigate wind shear and gust effects at the

moment of retro-rocket fire by ensuring that the angle between the retro-rocket thrust line is close to the local vertical. The TIRS consists of three small rocket motors, spaced at 120° increments around the back cover, with an inertial measurement unit to measure the tilt angle and the angular acceleration of the back cover. At the time of retro-rocket ignition, a variable thrust vector is applied to the back cover by the TIRS motors to counteract undesirable lateral motion (Fig. 18.15).

A modification to the TIRS system called DIMES was implemented when it was realized that the predicted winds at the first landing site could exceed the airbag capability. This allowed the system to reduce the steady-state drift velocity of the lander. Using a simple descent camera system, three pictures were taken at 5 s intervals during the terminal descent in order to calculate an approximate horizontal drift velocity. The TIRS could then be fired to modify the retro-rocket thrust line and thus counteract the drift, whilst also minimizing unwanted wind shear induced tilt.

18.4.4 Other Descent Systems

Gliding vehicles (e.g. Space Shuttle orbiter, Buran, X-38) use aerodynamic lift to control their descent rate and land on either a conventional aircraft undercarriage or skids. Such systems need a large area of flat, smooth terrain so are currently only appropriate for Earth return.

Autorotating helicopter-like blades have been considered for Mars landings [36] and have been studied for landing crewed capsules.

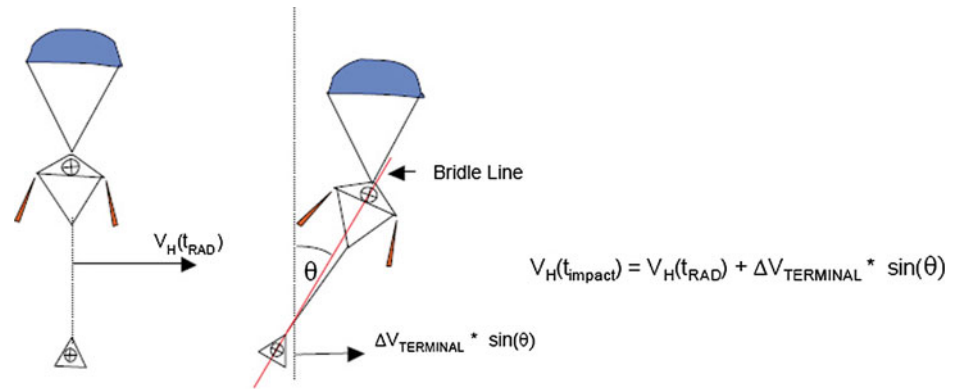
Balloons were used on the Vega Venus missions to maintain an altitude of 54 km above the surface of the planet and have been considered for Titan exploration [37].

18.5 Landing

The landing system operates from first impact with the ground until the vehicle comes to complete rest. During this time, it must absorb all residual energy and protect the payload.

In general, the more complex the terminal descent system the less complex the landing system needs to be. Landers with sophisticated controlled propulsive systems that are able to deliver the vehicle to the surface with small vertical and lateral velocities need only landing legs (Viking, Phoenix), crushable structures (ExoMars 2016), or the suspension of the rover (Mars Science Laboratory). Less sophisticated descent systems such as MPF, MER, and Beagle-2 needed airbags to protect the lander from the significantly increased range of landing velocities [38].

Fig. 18.15 Transverse impulse rocket system (TIRS) operation diagram



18.5.1 Landing Legs/Crushable Structures

Three deployable landing legs [33] provide a simple, reliable landing system. Landing legs have been used on past planetary landers, such as Surveyor, Apollo, Viking, and Phoenix. Rocks and slopes are the main hazard for this type of landing system. They are usually of inverted tripod or cantilever construction, with a crushable honeycomb shock absorber in one of the struts. Landers to date have had from 20 to 30 cm of ground clearance (after leg stroke for landing load attenuation) to avoid rock contact with the belly of the lander. Clearance is also required for the propulsion system. The terminal descent system can only be close to the ground for a very short time if it is to avoid severe regolith disturbance. For this reason, legged landers with integrated propulsion systems approach the ground at a relatively high speed (2.4 m/s) with a consequently increased susceptibility to slope-induced tip-over hazards. The ground is sensed by pads in the lander foot. The propulsion system is shut down immediately as first ground contact is made. Equipping the belly of the lander with a crushable structure is an even simpler approach.

18.5.2 Airbags

There are two types of airbag system: non-vented and vented. Non-vented systems act like bouncy balls and gradually bring the payload to rest while dissipating the impact energy. Vented systems seek to absorb the impact energy by compression of the airbag gas and immediately release the compressed gas to prevent a rebound. All of the extra-terrestrial landing systems that have flown to date (Table 18.3) have been non-vented systems.

Airbags were first used as planetary landing systems in the 1960s by Semyon Lavochkin (1900–1960) and Georgy Babakin (1914–1971) in Russia on some of the Luna series of Moon landers. The next use of airbag systems for extra-terrestrial missions was not until 1996 with the launch of the Russian Space Forces (RSF) Mars 96 and the NASA Mars

Pathfinder [39]. Mars 96 did not reach its destination owing to a failure of the final stage during launch, but Pathfinder was highly successful and airbags were therefore chosen for the NASA MER missions launched in 2003 [40]. Within Europe, an airbag landing system was developed for the Beagle-2 lander [41], which was flown as part of the 2003 ESA Mars Express mission, but with unknown results. An airbag system was originally baselined for the ESA Exo-Mars mission, with both non-vented and vented systems proceeding through to the phase-B stage before they were replaced by a propulsive descent and landing system.

Airbags have also been studied as landing or impact load attenuation systems for the Orion capsule [42] and the Kistler K1 rocket [43] to cushion the landing when returning to Earth. These airbags were of the vented type.

Non-vented airbags protect the lander by preventing contact with rocks or the ground and limiting the accelerations to which it is exposed. The airbag may take several bounces to come to rest, because to dissipate the kinetic energy of the vehicle it relies on thermodynamic losses during the compression and expansion of the gas in the airbag, friction between the airbag and the ground, and the work done during the stretching and relaxation of the airbag material. Due to the longer operational period of a non-vented airbag system and the possibility of multiple bounces and impacts with the ground before the payload comes to rest, the airbags must fully encompass the payload and maintain their structural integrity and gas-tightness for a much longer period than a vented system. Additionally, pressure loss due to cooling of the inflation gas and heat transfer to the environment is of greater concern for non-vented systems; both MPF and MER featured 'sustain charges' to manage the loss in pressure due to cooling. Although non-vented systems are heavy, this is mitigated by the greater simplicity of the system and the reduced sensitivity to attitude, lateral velocity, and slopes during impact. The uncontrolled nature of the landing sequence with a non-vented system means that it is almost impossible to guarantee the orientation of the payload when the system comes to rest. As a result, some form of self-righting mechanism is

Table 18.3 Historical use of airbag systems for extra-terrestrial space missions

Mission	Configuration	Payload mass	Airbag system mass
USSR Luna 9 and others (1960s)	Non-vented—spherical	99 kg	–
RSF Mars 96 (1996)	Non-vented—spherical	12 kg	–
NASA Mars Pathfinder (1996)	Non-vented—billiard rack	290 kg	99 kg
UK/ESA Beagle 2 (2003)	Non-vented—spherical	35 kg	15 kg
NASA Mars Exploration Rovers (2003)	Non-vented—billiard rack	415 kg	125 kg

required, either as part of the payload deployment or as a by-product of the airbag deflation or jettisoning process. Major design drivers for these systems are high lateral velocities, which increase the potential for abrasion damage and cause increased impact velocities in conjunction with slopes. Encounters with large pointed rocks are also a major source of failure due to puncture or tear damage.

Several geometries have been used. MPF and MER used a tetrahedral ‘billiard rack’ design, as seen in Fig. 18.16. After coming to rest the airbags were deflated, the lander righted, and the airbags retracted under the landing cage structure that protected the rover, as seen in Fig. 18.17. Beagle-2 used spherical airbags comprising three separable segments. When the bag stopped, the segments were to be released and separated by venting residual pressure, leaving the lander to fall to the surface from a height of approximately the radius of the airbag. A double toroid geometry has also been studied for small Mars landers.

Vented airbags are more complex than non-vented systems, requiring a degree of control to ensure optimal operation. Rather than relying on thermodynamic losses through successive compression and re-expansion of the inflation gas in order to dissipate energy, vented systems operate by compressing the inflation gas during the impact and then releasing the gas to the external environment before it can re-expand. This allows a vented system to stop the lander during a single impact event. The primary challenge for vented systems is to control the payload dynamics during and after venting, to prevent tip over. This challenge is exacerbated by high lateral velocities, slopes, and rocks. Vented systems are not very efficient at reducing horizontal or tangential velocities, and they rely on interaction of the lander with the terrain to remove the residual motion along the surface. It is during this phase that the lander is at the greatest risk of turning over. A vented airbag system that stops on the first bounce is tolerant to slow punctures and abrasions that cause a gradual loss in pressure, and offers lower system mass since only the underside of the payload needs protection.

Vented airbags are typically divided into several segments, with each segment separately vented. Control of venting based on acceleration sensors and laser range finders has been investigated. A particular challenge for Mars landing is that that low ambient pressure means that

the gas flow from the bag at venting is at sonic velocity, leading to rapid and complete venting of the segment. At terrestrial ambient pressure the gas vents at subsonic velocity and the segment rapidly reaches ambient pressure, with the remaining deflation being slow. Mars system venting therefore needs to be considerably more accurate in order to control the lander dynamics.

Studies for the ExoMars program showed that to achieve a high probability of landing success the lateral velocity for a vented system had to be controlled, necessitating a propulsive descent system which itself removed the need for an airbag system since the vertical velocity was also controlled and a crushable structure provided an adequate landing system.

18.5.3 Mid-Air Retrieval

The use of mid-air retrieval avoids a landing system entirely for Earth return missions. For Genesis it was planned that a helicopter would snag the parafoil descent system in mid-air and then fly the probe to the ground [15]. It should however be noted that planetary protection best-practice typically requires a high probability of any sample canister remaining intact, and hence sealed, in the event of a total failure of the EDL system. This requirement can drive an Earth return to adopt this as the default mission baseline.

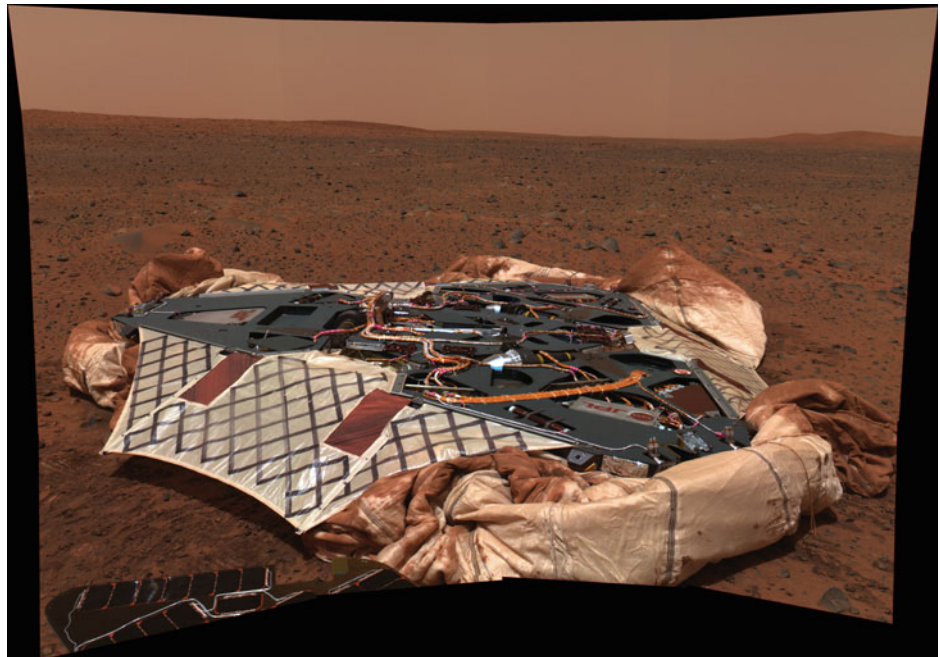
18.6 Modeling and Simulation

Entry, descent and landing systems for space vehicles rely heavily on modeling and simulation throughout the design, development, and qualification process. During the design phase, computer modeling allows the EDL sequence to be optimized and the system components to be sized. During development and qualification, modeling allows appropriate test conditions to be derived and test results to be cross-referenced against mission conditions. Simulations are far more important for space missions than terrestrial descent and landing systems, since it is generally impossible to test hardware in conditions which are completely representative of the mission environments (e.g. gravity, atmosphere density, and composition).

Fig. 18.16 MER airbags tetrahedral ‘billiard rack’ design.
Image NASA/JPL/Cornell



Fig. 18.17 Image mosaic taken by the panoramic camera on-board the Mars Exploration Rover Spirit showing the rover’s landing site, the Columbia Memorial Station, at Gusev Crater, Mars. The airbags are seen underneath the rover platform. *Image NASA/JPL/Cornell*



18.6.1 Typical Simulation Tools

A number of simulation tools are used in the design of an EDL system. Some are specific to EDL design, others are more general. The specialized tools include the following.

- Trajectory simulation tools are used to simulate the entire time from atmospheric entry to landing. They vary from simple, three degrees of freedom (3dof), point-mass tools to complex multi-body, six degrees of freedom (6dof) models that include interactions between the bodies. These are described in more detail later in this section.

Examples include the NASA POST2 [44] tool and the ESA EAGLE software [45]. Most aerospace companies also have customized in-house tools with similar capabilities.

- Entry thermodynamics tools model the performance of ablative thermal protection systems throughout the entry phase. These models predict the heat flux onto the thermal protection system (both radiative and convective) and the response of the thermal protection material. They range in fidelity from simple empirical models to those using quantum chemistry to model the hypersonic flow

physics, conventional chemistry to model the dissociation of the TPS, and CFD to model the flow of the ablation products from the aeroshell.

- Inflation models are used to predict the inflation process of parachutes. These may be low-fidelity models built into trajectory simulation tools for system sizing or high-fidelity models that facilitate detailed stressing of the parachute and payload.

18.6.2 Development Process

Simulations are used in different ways throughout the development process of any EDL system. The fidelity of the simulation will usually increase as the design matures.

During the conceptual design phase, the objective is to size the overall EDL system within the mission constraints (e.g. allowable acceleration, entry angles, and landing site altitudes). The initial analysis is generally performed using a model that incorporates empirical models of heat flux, thermal protection system performance, and aerodynamics. This allows design options to be traded (e.g. the number of parachutes and the type of landing system), for the key components to be sized, and for the initial mass estimates to be made.

Once a design has been chosen, its performance can be assessed throughout the operational envelope using the same trajectory tool in a Monte-Carlo mode. This allows thousands of simulations to be run while varying initial states, environmental conditions, and system parameters, in order to assess the robustness of the chosen design and choose the design parameters.

The detailed design of the entry, descent, and landing subsystems will make use of specific tools to predict the performance at each stage. Examples are parachute deployment and inflation and airbag operation. Multi-body, 6dof tools are frequently used to predict the relative motion of the components that are released during the sequence to ensure that none of these items will subsequently interfere with the entry vehicle.

During the development phase, test data are used to enhance the fidelity of the simulations. Since the mission environment cannot generally be replicated on Earth, the simulations are run using both Earth test conditions and mission conditions. They are validated against test conditions on Earth, and then used to extrapolate and calculate the likely performance under mission conditions.

18.6.3 The Anatomy of Simulations

Simulations bring together models of physical processes (either empirical or physics-based) and environments

employing numerical methods for their solution. These are frequently produced as standard libraries or building blocks which may subsequently be used in different software and for different purposes; for example an atmosphere model for a planet may be used in dynamic simulations, parachute inflation simulations, or landing predictions.

Models may be produced for differing levels of fidelity, ranging from simple empirical models to detailed physical models. The former will require little computing power and may be used for initial design and iterative tasks such as Monte-Carlo simulations, whereas the latter may require much greater resources and be suitable for high-fidelity predictions of the final mission design.

The key building blocks of simulations are described in the following sections.

18.6.4 Reference Frames

Since EDL sequences occur on (nearly) spherical, rotating planets and moons, the choice of simulation reference frame is important. The distances covered during entry and descent are generally sufficiently great that a flat planet model cannot be used; however, it may be useful for terminal descent and landing simulations. For a three-dimensional planet model, the reference frame may be inertial (fixed with respect to space) or rotating (fixed with respect to the planet). Both solutions have been used and have different challenges. If a rotating reference frame is used, care must be taken to consider Coriolis effects.

18.6.5 Atmosphere Models

Wherever a target body possesses an atmosphere, this is inevitably utilized during the EDL process. In order to design the mission, atmosphere models are required. For initial mission design, these can be as simple as a single density profile with altitude; for detailed design, additional parameters and their variation must be considered.

The density profile is the most important atmospheric property for the design of an EDL system since it, along with airspeed, determines the aerodynamic force on an entry vehicle. For some destination, such as Earth, the atmosphere is sufficiently dense and homogenous that the majority of mission design can take place using a single density profile; for others, such as Mars, the spatial, diurnal, and seasonal variations in the atmosphere are significant and must be considered from the earliest stages of mission design.

The temperature profile is important to define the speed of sound in the atmosphere, and thus the Mach number during the EDL operation. It can also be important for the choice of materials for the descent and landing systems.

The thermodynamic and transport properties of the atmosphere allow determination of the flow characteristics around the vehicle by reference to dimensionless quantities such as the Mach, Reynolds, and Strouhal numbers of the flow.

Perturbation models are necessary for detailed design of the descent and landing systems. The principal perturbation of interest is wind. The magnitude of the wind determines the ground speed of the payload during the terminal descent and thereby the lateral velocity requirements for the landing system, since any aerodynamic system will operate in a wind-relative frame. The rate of change of wind velocity with altitude (wind shear) and time (wind gusts) affects the stability of a payload under a parachute and the ability of a powered descent system to control a vehicle. Knowledge of the wind characteristics may be used to optimize the system geometry and to set requirements for key components such as landing radar.

Perturbations may be generated on several scales. They are generally split into two classes: large scale, with length scales of several kilometers caused, for instance, by atmospheric gravity waves; and small scale, with length scales of tens of meters and less, caused by atmospheric turbulence. Typical engineering atmosphere models predict randomized perturbations based on global conditions, and are sufficiently accurate for initial mission design, but local climate modeling (meso-scale modeling) may be required for the intended landing site if the effects of wind are critical.

When using atmosphere models as part of a dynamic simulation, care must be taken over the integration of the atmosphere model with the terrain model (in order to determine when the vehicle has reached terminal conditions). The atmospheric datum (zero altitude) rarely corresponds to ground level. Usually it relates to a surface with a given gravitational potential; however, for some planets several 'standard' reference surfaces have been defined.

Engineering models for most potential destinations have been produced. NASA Marshall Space Flight Center has produced models in the GRAM (Global Reference Atmosphere Model) for Venus, Earth, Mars, Neptune, and Titan, all of which feature atmospheres with random perturbations. Other organizations have also produced a selection of atmosphere models.

The atmosphere of Venus has a surface density of 66 kg/m³ and a pressure of 95 bars. Standard atmosphere profiles such as the Venus International Reference Atmosphere (VIRA) [46] may be used for initial mission design. VenusGRAM [47], which includes perturbations, may be considered for more realistic mission design.

The atmosphere of Earth has obviously been studied in far greater detail than those of the other planets and many engineering models are available. For initial studies, the ICAO standard atmosphere [48] is usually sufficient. Extreme atmosphere profiles have been documented by the

US Department of Defense [49] up to an altitude of 80 km. For profiles to the upper limits of the atmosphere, the NRL MSISE-00 [50] model predicts density and temperature as a function of latitude, longitude, altitude, and time. It is freely available but does not include perturbations. The EarthGRAM [51] model may be used if random perturbations are required.

Due to the tenuous nature of the atmosphere of Mars and its temporal variability, single standard profiles are not suitable for mission design. Two global engineering models are in common use: MarsGRAM and the European Mars Climate Database (EMCD). The latter constructs conditions using databases that include mean conditions and variability derived from full-planet climatic simulations.

For the gas giants (Jupiter, Saturn, Uranus and Neptune), the only engineering model known to the authors is NeptuneGRAM. For missions to these planets, an early task may be the construction of such an atmosphere if required, however this would be a mission specific task.

The final target of interest, and the only moon in the solar system with a viable atmosphere, is Saturn's moon, Titan. Several engineering models have been produced for Titan including those by Lellouche and Hunten [52], Yelle [53], and TitanGRAM.

A good summary of atmosphere engineering models has been produced by the AIAA [54].

18.6.6 Gravity Models

Very accurate gravitational models are required in order to predict orbital evolution accurately. For entry simulations however, the aerodynamic forces on the entry body dwarf all but the lowest order terms. Furthermore, the errors due to uncertainties in atmosphere and aerodynamics knowledge will far exceed the effect of the higher order gravitational terms. Generally, no more than the second order gravitational term need be considered.

18.6.7 Aerodynamics

In order to predict the trajectory of an entry vehicle through an atmosphere, the aerodynamics of the vehicle must be known. As discussed in Chap. 5, the flight of an entry vehicle will encompass all flow regimes from free molecular to continuum and from hypersonic to subsonic. For mission design, good estimates of the aerodynamics of the vehicle must be used. Many papers describing the aerodynamics of past systems have been published and these are useful for initial system design. As the mission design matures, it is often necessary to supplement the literature values using either CFD or testing.

18.6.8 Dynamics

Two classes of simulation are used during mission design and validation: three degrees of freedom (3dof) and six degrees of freedom (6dof).

Three degrees of freedom simulations model the object(s) under consideration in terms of point masses. These are only useful for ballistic entries (i.e. non-lifting), although the addition of a fourth degree of freedom for simple lifting bodies is possible. These are useful for initial mission design, and may indeed be used for trajectory modeling throughout mission preparation for non-lifting systems. Their simplicity allows trajectories to be calculated very quickly.

Six degrees of freedom simulations model the position and orientation of each body under consideration (i.e. 3 element linear position and 3 element rotational position). Each body (e.g. entry vehicle, released heat shield, parachutes, and so forth) is modeled as a separate body with its own dynamics and aerodynamics. Bodies may be connected (e.g. parachutes) using elastic members which enable the force transfer to be represented and quantified. These simulations facilitate detailed modeling of stability during entry, the dynamics of separation and re-contact events, the forces in parachute risers, and many other aspects of the performance of a system. However, they are complex and time-consuming to configure. While a 3dof simulation needs only mass and aerodynamic drag for each object, a 6dof simulation requires masses, inertias, full aerodynamics, center of gravity positions and attachment positions for each object.

18.6.9 Simulation Features

The ability to perform Monte-Carlo runs is essential in modeling EDL systems. There are always uncertainties in the properties of the target body and indeed those of the spacecraft. It is not generally worthwhile, or practical, to design a system for all combinations of worst-case conditions, so a Monte-Carlo approach is usually adopted to define design cases. This can involve randomizing the values of tens of properties for a typical system. The simulation software should be capable of performing these randomized simulations and analyzing the output files in order to produce descriptive statistics. This may be achieved either as part of the main software or using a driver program to run the simulation software in batch mode.

References

1. Milos FS, Chen YK, Squire TH, Brewer RA. Analysis of Galileo Probe Heatshield Ablation and Temperature Data. *Journal of Spacecraft and Rockets*. 1999; 36(3).
2. Thurman SW. Surveyor Spacecraft Automatic Landing System. In 27th Annual AAS Guidance and Control Conference; 2004; Breckenridge, Colorado.
3. Spencer DA, Blanchard RC, Braun RD, Kallemeyn PH, Thurman SW. Mars Pathfinder Entry, Descent and Landing Reconstruction. *Journal of Spacecraft and Rockets*. 1999; Vol 36 No 3.
4. Crowder RS, Moote JD. Apollo Entry Aerodynamics. *Journal of Spacecraft and Rockets*. 1969; Vol 6 No 3.
5. NASA. Press Kit - Apollo 7. Press Kit. NASA; 1968. Report No.: 68-168 K.
6. Sutton K, Graves RA. A General Stagnation Point Convective Heating Equation for Arbitrary Gas Mixtures. Technical Report. Washington D.C.: NASA; 1971. Report No.: NASA TR-376.
7. Tauber ME, and Sutton K. Stagnation-Point Radiative Heating Relations for Earth and Mars Entries. *Journal of Spacecraft and Rockets*. 1991.; vol. 28, no. 1, pp. 40–42.
8. Laub B, Venkatapathy E. Thermal Protection System Technology and Facility Needs for Demanding Future Planetary Missions. In International Workshop on Planetary Probe Atmospheric Entry and Descent Trajectory Analysis and Science; 2003; Lisbon, Portugal.
9. Marraffa L, Kassing D, Baglioni P, Wilde D, Walther S, Pitchkhadze K, et al. Inflatable Re-Entry Technologies: Flight Demonstration and Future Prospects. *ESA Bulletin*. 2000 Aug;(103).
10. Brown GJ, Epp C, Graves C, Lingard JS, Darley MG, Jordan K. Hypercone Inflatable Supersonic Decelerator. In 17th AIAA Aerodynamic Decelerator Systems Technology Conference; 2003; Monterey, CA: AIAA 2003-2167.
11. Cruz JR, Lingard JS. Aerodynamic Decelerators for Planetary Exploration: Past, Present and Future. In AIAA Guidance, Navigation, and Control Conference and Exhibit; 2006: AIAA 2006-6792.
12. Vincze J. Gemini Spacecraft Parachute Landing System. Technical Note. Washington D.C.: NASA; 1966. Report No.: NASA TN D-3496.
13. Knacke TW. The Apollo Parachute Landing System. In AIAA Second Aerodynamic Decelerator Systems Conference; 1968; El Centro, California. p. TP-131.
14. Bienstock BJ. Pioneer Venus and Galileo Entry Probe Heritage. In International Workshop on Planetary Probe Atmospheric Entry and Descent Trajectory Analysis and Science; 2003; Lisbon, Portugal.
15. Smith J, Witkowski A, P. W. Parafoil Recovery Subsystem for the Genesis Space Return Capsule. In 16th AIAA Aerodynamic Decelerator Systems Technology Conference; 2001; Boston MA: AIAA 2001-2017.
16. Stein JM. Parachute Testing for the NASA X-38 Crew Return Vehicle. In 36th Annual International Symposium; 2005; Fort Worth, TX; United States.
17. Ewing EG, Bixby HW, Knacke TW. Recovery Systems Design Guide. Technical Report. Dayton, OH: Air Force Flight Dynamics Laboratory; 1978. Report No.: AFFDL-TR-78-151.
18. Mitcheltree R, Bruno R, Slimko E, Baffes C, Konefat E, Witkowski A. High Altitude Test Program for a Mars Subsonic Parachute. In 18th AIAA Aerodynamic Decelerator Systems Technology Conference; 2005; Munich: AIAA 2005-1659.
19. Fallon II EJ. System Design Overview of the Mars Pathfinder Parachute Decelerator System. In 14th AIAA Aerodynamic Decelerator Systems Technology Conference; 1997; San Francisco, CA: AIAA 97-1511.
20. Witkowski A, Kandis M, Bruno R, and Cruz JR. Mars Exploration Rover Parachute System Performance. In 18th AIAA Aerodynamic Decelerator Systems Technology Conference and Seminar; 2005; Munich: AIAA 2005-1605.

21. Lingard JS, Underwood JC. The Effects of Low Density Atmospheres on the Aerodynamic Coefficients of Parachutes. In 13th AIAA Aerodynamic Decelerator Systems Technology Conference; 1995; Clearwater Beach, FL: AIAA 95-1556.
22. Lingard JS, Darley MG, Underwood JC. Simulations of Mars Supersonic Parachute Performance and Dynamics. In 19th AIAA Aerodynamic Decelerator Systems Technology Conference; 2007; Williamsburg, VA: AIAA 2007-2507.
23. Lingard JS. A Semi-Empirical Theory to Predict the Load-Time History of an Inflating Parachute. In 8th AIAA Aerodynamic Decelerator and Balloon Technology Conference; 1984; Hyannis, MA: AIAA 84-0814.
24. Doherr K. Extended Parachute Opening Shock Estimation. In 17th AIAA Aerodynamic Decelerator Systems Technology Conference; 2003; Monterey, CA: AIAA 2003-2173.
25. Pflanz E. Zur Bestimmung der Verzögerungskräfte bei Entfaltung von Lastenfallschirmen.; 1942. Report No.: ZWB FB 1704.
26. McVey DF, Wolf DF. Analysis of Deployment and Inflation of Large Ribbon Parachutes. *Journal of Aircraft*. 1974; 11 No 2.
27. Ludtke WP. Notes on Genexix Parachute Opening Force Analysis.; 1986. Report No.: NSWC TR-86-142.
28. Murrow HN, McFall Jr JC. Some Test Results from the NASA Planetary Entry Parachute Program. *Journal of Spacecraft and Rockets*. 1969; 6 No 5.
29. Underwood JC. Development Testing of Disk-Gap-Band Parachutes for the Huygens Probe. In 13th AIAA Aerodynamic Decelerator Systems Technology Conference; 1995; Clearwater Beach, FL: AIAA 95-1549.
30. Witkowski A. The Stardust Sample Return Capsule Parachute Recovery System. In 15th CEAS/AIAA Aerodynamic Decelerator Systems Technology Conference; 1999; Toulouse, France: AIAA 99-1741.
31. Lingard JS. The Performance and Design of Ram-Air Gliding Parachutes. Technical Report. Farnborough, UK: Royal Aircraft Establishment; 1981. Report No.: RAE TR-81103.
32. Pleasants JE. Parachute Mortar Design. *Journal of Spacecraft and Rockets*.; Vol. 11, p.246.
33. Manning RM, and Adler M. Landing on Mars. In AIAA Space 2005 Conference; 2005; Long Beach, CA: AIAA 2005-6742.
34. Steltzner AD, Burkhart PD, Chen A, Comeaux KA, Guernsey CS, Kipp DM, et al. Mars Science Laboratory Entry, Descent and Landing System Overview. In 7th International Planetary Probe Workshop; 2010; Barcelona, Spain.
35. Steltzner A, Desai P, Lee W, Bruno R. The Mars Exploration Rovers entry descent and landing and the use of aerodynamic decelerators. In 17th AIAA Aerodynamic Decelerator Systems Conference; 2003; Monterey, CA: AIAA 2003-2125.
36. Lutz T, Westerholt U, Noeding P, Ransom S, Köhler J. Application of Auto-Rotation for Entry, Descent, and Landing on Mars. In 7th International Planetary Probe Workshop; 2010; Barcelona, Spain.
37. Lorenz RD. A Review of Balloon Concepts for Titan. *Journal of the British Interplanetary Society*. 2008; 61.
38. Rivellini T. The Challenges of Landing on Mars. *Bridge(USPS 551-240)*, National Academy of Engineering. 2004; 34(4).
39. D. C, C. S, M. G. Development and Evaluation of the Mars Pathfinder Inflatable Airbag Landing System. *Acta Astronautica*. 2002 May; 50(10).
40. Stein J, Sandy C. Recent Developments in Inflatable Airbag Impact Attenuation Systems for Mars Exploration. In 2nd International Symposium Atmospheric Reentry Vehicles and Systems; 2003; Arcachon, France: AAAF-61.
41. Huxley-Reynard CS. An Airbag System for the Beagle2 Mars Probe. In 16th AIAA Aerodynamic Decelerator Systems Technology Conference; 2001; Boston, MA: AIAA 2001-2046.
42. Tutt B, Sandy C, Corliss J. Status of the Development of an Airbag Landing System for the Orion Crew Module. In 20th AIAA Aerodynamic Decelerator Systems Technology Conference; 2009; Seattle, WA: AIAA 2009-2923.
43. Gardinier DJ, Taylor AP. Design and Testing of the K-1 Reusable launch Vehicle Landing System Airbags. In 15th AIAA Aerodynamic Decelerator Systems Technology Conference; 1999; Toulouse, France: AIAA 97-1757.
44. Murri DG. Simulation Framework for Rapid Entry, Descent and Landing (EDL) Analysis. Hampton, VA: NASA Langley Research Center; 2010. Report No.: NASA/TM-2010-216867.
45. St. John-Olcayto E, Johns G, Pidgeon A, Philippe C. EAGLE: An Extensible, End to End Simulation and Evaluation Framework for Planetary E/DLS. In International Planetary Probe Workshop 2010: IPPW7; 2010; Barcelona.
46. Kliore A. The Venus international reference atmosphere. *Advances in Space Research*. 1986.
47. Justus HL, Justus CG, Keller VW. Global Reference Atmosphere Models including Thermospheres for Mars, Venus and Earth. In AIAA/AAS Astrodynamics Specialist Conference and Exhibit; 2006; Keystone, CO.
48. International Organisation for Standardization. Standard Atmosphere. Standard.; 1975. Report No.: ISO 2533:1975.
49. Department of Defense. Global Climatic Data for Developing Military Products. Department of Defense Handbook.; 1997. Report No.: MIL-HDBK-310.
50. Picone JM, Hedin AE, Drob DP, Aikin AC. NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *J. Geophys Res*. 220; 107(A12).
51. Justus CG, Leslie FW. The NASA MSFC Earth Global Reference Atmospheric Model - 2007 Version.; 2008. Report No.: NASA/TM-2008-215581.
52. Lellouche E, Hunten DM. Titan Atmosphere Engineering Model.; 1987. Report No.: ESLAB 87-199.
53. Yelle RV, Strobell DF, Lellouch E, Gautier D. Engineering Models for Titan's Atmosphere.; 1994.
54. AIAA. Guide to Reference and Standard Atmosphere Modes. Guide.; 2010. Report No.: AIAA G-003C-2010.

Kazuya Yoshida, Dragomir Nenchev, Genya Ishigami
and Yuichi Tsumaki

This chapter discusses robotics technology for space missions. First, a general definition of a robot and an overview of the historical development of space robots are provided. Then technical details of orbital space robots, planetary robots, and telerobotics are given in the subsequent sections.

The term ‘robot’ comes from the word ‘robota’, which means serf labor or hard work in the Slavic languages (Czech, Slovak and Polish). It was largely introduced to the public by the Czech writer Karel Čapek (1890–1938) in his play *R.U.R.* (*Rossum’s Universal Robots*), which was premiered in 1920. In this play, the robots are described as artificial creatures, or androids, which can be mistaken for humans.

Today, the word robot is used for an intelligent machine or artificial agent that can exhibit interactive behavior with its environment or a human in a coordinated manner. Although humanoids, or human-looking robots, have attracted public attention, the typical robots used in industry are automated or programmable handling devices that do not necessarily look like humans. Actually, many such industrial robots are successfully working in the mass-production lines of industrial factories, conducting repetitive

tasks such as welding or assembling motor vehicles. However, the majority of research efforts now involve robots that can work outside the factory, such as in offices, homes and hospitals, or in outdoor fields or outer space (space robot, the focus of this chapter), or even in inner space (medical robots, which can work inside the human body). Robotics is a discipline involving system integration, which forms the basis for most of our knowledge of many different subject areas including mechanics, electronics, computer technology, and bioengineering, along with various topics in human sciences, such as anthropology and sociology.

Autonomy is a key issue in robotics, and at a primitive level, any non-crewed spacecraft that is under automated sequence control may be referred to as a robotic satellite. However, when the term space robot is used it implies a more capable mechanical system that can facilitate manipulation, assembly, or service tasks in orbit as an assistant to astronauts, or can extend the areas and abilities of exploration on remote planets as a surrogate for human explorers.

The key issues in space robotics are characterized as follows

- **Manipulation**—Although manipulation is a basic technology in robotics, the microgravity of the orbital environment requires special attention to the motion dynamics of the manipulator arms and the objects being handled. The reaction dynamics that affect the base body, impact dynamics when the robotic hand contacts an object to be handled, and vibration dynamics due to structural flexibility are included in this issue. Technical details of the manipulator control in the microgravity environment are elaborated in [Sect. 19.2](#).
- **Mobility**—Locomotion is particularly important in exploration robots (rovers) that travel on the surface of a moon or planet. These surfaces are natural and rough, and thus challenging to traverse. Sensing and perception, traction mechanics, and vehicle dynamics, control and navigation are all mobile robotics technologies that must

K. Yoshida (✉)

The Space Robotics Laboratory, Department of Aerospace Engineering, Tohoku University, Sendai, Japan
e-mail: yoshida@astro.mech.tohoku.ac.jp

D. Nenchev

Department of Mechanical Systems Engineering, Tokyo City University, Tokyo, Japan

G. Ishigami

Department of Mechanical Engineering, Keio University, Yokohama, Japan

Y. Tsumaki

Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan

be demonstrated in a natural untouched environment. Technical details of the surface mobility systems are elaborated in Sect. 19.3.

- *Teleoperation and Autonomy*—There is non-negligible time delay between a robotic system in space and a human operator in an operation room on Earth. In early orbital robotics demonstrations, this latency was typically a few seconds, but can be several tens of minutes, or even hours for planetary missions. Telerobotics technology is therefore indispensable in space exploration, and the introduction of autonomy is a reasonable consequence. Technical details of the telerobotics are elaborated in Sect. 19.4.
- *Extreme Environments*—In addition to the microgravity environment, which affects the motion dynamics of a robot, there are many other issues related to extreme space environments that are challenging and must be solved to enable practical engineering applications. Such issues include extremely high or low temperatures, high vacuum or high pressure, corrosive atmospheres, ionizing radiation, and very fine dust, and were discussed in detail in Chap. 3.
- *Versatility*—This is the ultimate goal when designing and developing a robot, and is especially highlighted in space applications. Due to the nature of space missions, once launched into space, a robot must perform all of its tasks by itself using its own resources. A space robot, therefore, should be adaptable to the extreme space environments mentioned above and possess the versatility to handle many different situations and scenarios, including contingent ones that arise unexpectedly.

19.1 Overview of the Historical Development of Space Robots

19.1.1 Orbital Space Robots

The first robotic manipulator arm used in the orbital environment was the Space Shuttle Remote Manipulator System (SRMS). It was successfully demonstrated in the STS-2 mission in 1981. This success opened a new era of orbital robotics and inspired numerous mission concepts.

A long-term goal that has been discussed extensively since the early 1980s is the application of a robotic free-flyer or free-flying space robot to the rescue and servicing of malfunctioning spacecraft (for example, the ARAMIS report [1]). In later years, crewed service missions were conducted for the capture-repair-deploy procedure of a malfunctioning satellite (Intelsat 603 by STS-49, for example) and for the maintenance of the Hubble Space Telescope (STS-61, -82, -103, -109 and -125). In each of these examples, the Space Shuttle, a crewed spacecraft with



Fig. 19.1 Space shuttle remote manipulator system (SRMS) used as a platform for an astronaut's extravehicular activity in the Shuttle cargo bay. Image NASA/CSA

dedicated maneuverability, was used.¹ In contrast, non-crewed servicing missions have not yet become operational. Although there have been several demonstration flights such as ETS-VII and Orbital Express, the practical technologies for non-crewed satellite servicing missions await the outcomes of future challenges.

19.1.1.1 Space Shuttle Remote Manipulator System

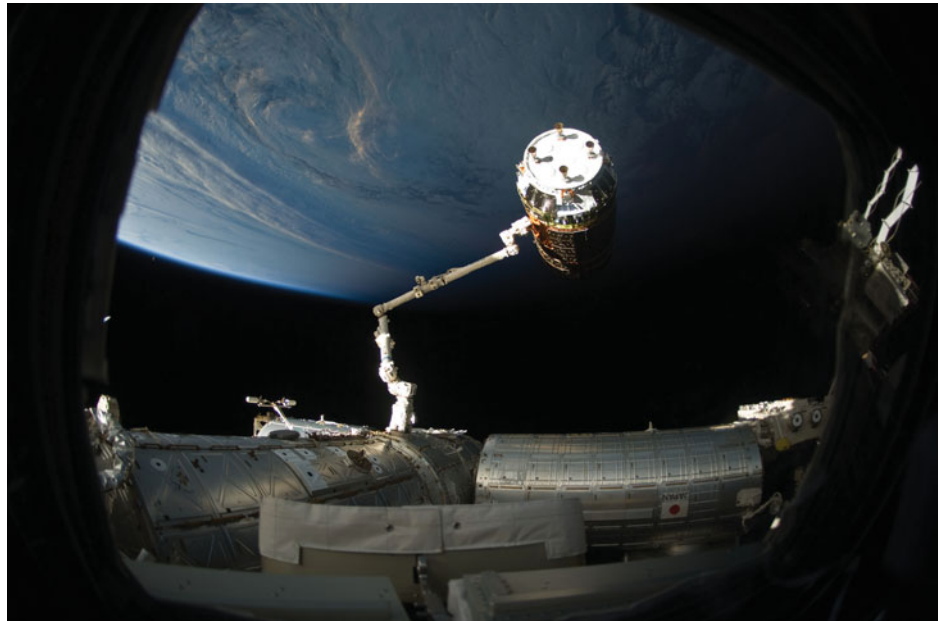
On-board the Space Shuttle, the Shuttle Remote Manipulator System (SRMS), or Canadarm, was a mechanical arm that handled a payload from the payload bay of the Space Shuttle orbiter. It could also grapple a free-flying payload and maneuver it into the payload bay. The SRMS was first used on mission STS-2, launched in 1981. It was used more than 100 times during subsequent missions, performing payload deployment and retrieval, as well as assisting in human extra vehicular activities (EVA) or space walks.² Servicing and maintenance missions to the Hubble Space Telescope and construction tasks for the International Space Station were also carried out by the cooperative use of the SRMS in human EVAs.

The SRMS arm was 15 m long and had six degrees of freedom (DOF), comprising shoulder yaw and pitch joints, an elbow pitch joint, and wrist pitch, yaw, and roll joints. Attached to the end of the arm was a special gripper system

¹ A robotic maintenance mission of the Hubble Space Telescope was seriously studied after the Space Shuttle *Columbia* accident (STS-107), but it was finally conducted as a crewed mission by STS-125.

² Five arms were built in total but one was destroyed in the *Challenger* accident in 1986.

Fig. 19.2 Space station remote manipulator system (SSRMS) grapples the Japan Aerospace Exploration Agency (JAXA) H-II transfer vehicle (HTV) prior to berthing it to the station. *Image NASA*



called the Standard End Effector (SEE), which was designed to grapple a pole-like fixture (GF) attached to the payload. By attaching a foothold at the end point, the arm could serve as a mobile platform for an astronaut's EVA, see Fig. 19.1.

19.1.1.2 International Space Station Mounted Robot Manipulator Systems

The International Space Station (ISS) is the largest international space project to-date, with 15 countries making significant cooperative contributions. The ISS is an outpost for the human presence in space, as well as a flying laboratory with substantial facilities for science and engineering research. To facilitate various activities on the station, there are several robotic systems.

The Space Station Remote Manipulator System (SSRMS), or Canadarm2, see Fig. 19.2, is an extended version of SRMS for use on the ISS. Launched in 2001 by STS-100 (ISS assembly flight 6A), the SSRMS has played a key role in the construction and maintenance of the ISS both by assisting astronauts during EVAs and in the use of the SRMS to hand over a payload from the Shuttle to the SSRMS. As for extensive capability, the SSRMS was designed as a symmetric seven-DOF arm with offset joints to enable it to be folded in half in the stored configuration and it provides system redundancy in operation. Its total length is 17.6 m when fully extended. Latching End Effectors are attached to both ends, through which power, data, and video can be transmitted to and from the arm. The SSRMS is self-relocatable using an inchworm-like movement with alternate grapples of Power Data Grapple Fixtures (PDGF), which are installed all over the station's

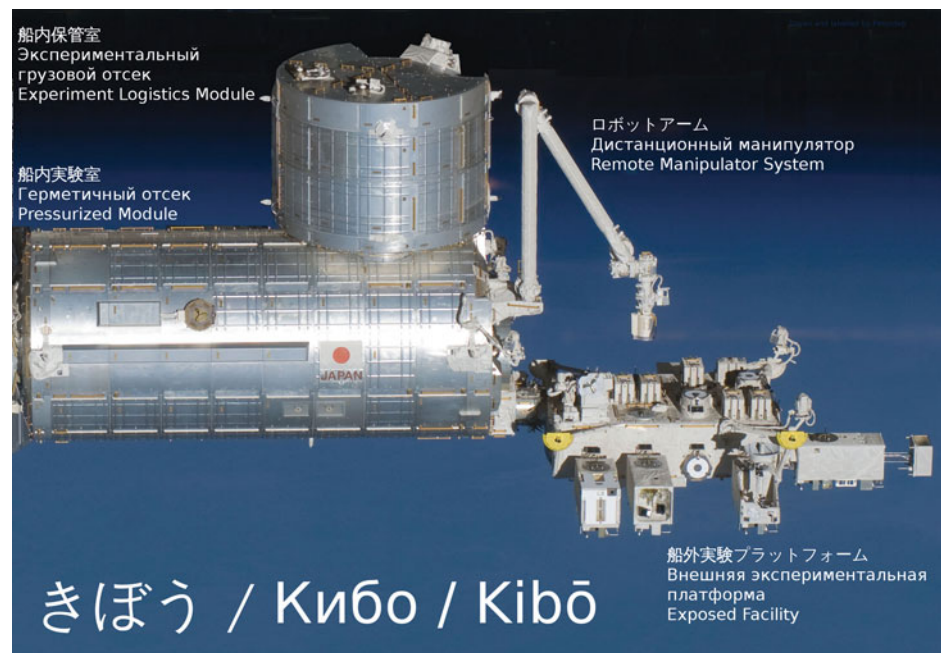
exterior surfaces to provide the power, data, and video, as well as a footholds. As another mobility aid to allow the SSRMS to cover wider areas of the ISS, Mobile Base System (MBS) was added in 2002 by STS-111 (ISS assembly flight UF-2). The MBS provides lateral mobility as it traverses the rails on the main trusses.

The Special Purpose Dexterous Manipulator (SPDM), or Dextre, which was attached at the end of the SSRM in 2008 by STS-123 (ISS assembly flight 1J/A), is a capable mini-arm system that facilitates the delicate assembly tasks currently handled by astronauts during EVAs. The SPDM is a dual-arm manipulator system, where each manipulator has seven DOFs and is mounted on a one-DOF body joint. Each arm has a special tool mechanism dedicated to the handling of standardized orbital replacement units (ORU) [2].

The Japan Space Exploration Agency (JAXA) also provided orbital assets including a robotic manipulator system for the ISS. The Japanese Experiment Module (JEM), which is also known by the nickname Kibo is composed of a pressurized module, exposed facility, experiment logistics module, and remote manipulator system (JEMRMS), see Fig. 19.3. These modules were developed by JAXA and successfully incorporated into the ISS by STS-123, 124 and 127 in 2008–2009.

The JEMRMS comprises two components: the main arm, which is a 9.9-m-long, six-DOF arm, and the small fine arm, which is a 1.9-m-long, six-DOF arm. Unlike the SSRMS, the main arm does not have self-relocation capability. Since its installation, the arm has been used to handle and relocate components for the experiments and observations performed in the exposed facility.

Fig. 19.3 The Japan Space Exploration Agency (JAXA) module, Kibo in orbit; other modules of the International Space Station have been removed through image manipulation. *Image creative commons*



19.1.1.3 ROTEX and ROKVISS

The robot technology experiment, ROTEX, which was developed by the German Aerospace Agency (DLR), is one of the historical milestones of robotics technology in space [3]. A multisensory robotic arm was flown on the Space Shuttle *Columbia* (STS-55) in 1993. Although the robot was confined to a work cell on the Shuttle, several key technologies were successfully tested, including those for a multisensory gripper, teleoperation from the ground and by the astronauts, shared autonomy, and time-delay compensation by the use of a predictive graphic display.

DLR also developed a two-joint manipulator system called ROKVISS, which was installed on the exterior of the Russian Service Module of the ISS in January 2005. The aim of ROKVISS was the in-flight verification of highly integrated modular lightweight robotic joints, as well as that of control technology, such as high-level system autonomy and force feedback-based teleoperation. The teleoperation experiments were conducted from the ground station via a direct radio link [4]. After 6 years of experiments in space, the ROKVISS flight hardware was brought back to Earth by a Soyuz return capsule.

19.1.1.4 Orbital Express and ETS-VII: 'Orihime' and 'Hikoboshi'

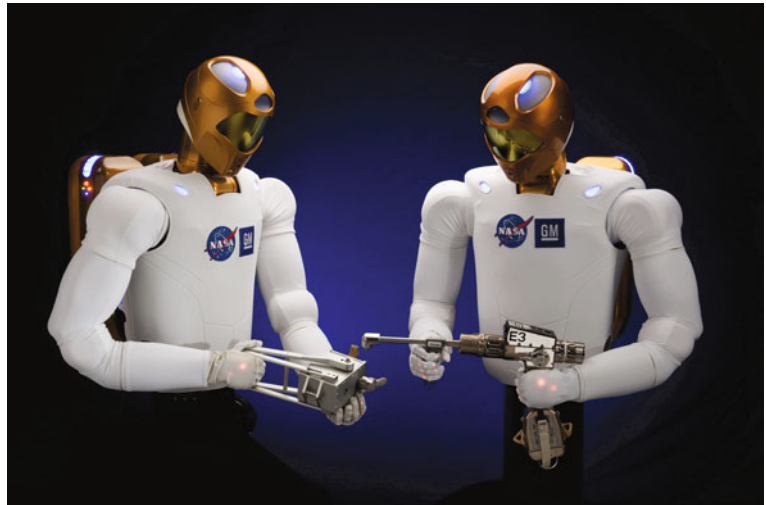
Japanese Engineering Test Satellite VII was another historical milestone in the development of robotics technology in space, particularly in the area of satellite servicing. ETS-VII was developed and launched by the National Space Development Agency of Japan (NASDA, currently JAXA) in November 1997. Numerous experiments were

successfully conducted using a 2-m-long, six-DOF manipulator arm mounted on its carrier satellite.

The mission objective of ETS-VII was to test free-flying robotics technology and to demonstrate its utility in orbital operation and servicing tasks. The mission consisted of two subtasks: autonomous rendezvous/docking (RVD) and numerous robot experiments (RBT). For the RVD experiments, the spacecraft was separated into two sub-satellites in orbit, one called 'Orihime', which behaved as a target, and the other called 'Hikoboshi', which acted as a chaser. The robot experiments included: (1) teleoperation from the ground with a time delay of 5–7 s. (2) Robotic servicing task demonstrations such as orbital replacement unit (ORU) exchange, fuel transfer between the satellite and the ORU, and deployment of space structures; (3) dynamically coordinated control between the manipulator reaction and the satellite attitude response; and (4) the capture and berthing of a target satellite, all of which were conducted successfully [5, 6].

Ten years after ETS-VII, a similar orbital demonstration was conducted under the Orbital Express Space Operations Architecture program by the Defense Advanced Research Projects Agency (DARPA) in the United States. The system consisted of the Autonomous Space Transport Robotic Operations (ASTRO) vehicle, developed by Boeing Integrated Defense Systems, and a prototype modular next-generation serviceable satellite, NextSat, developed by Ball Aerospace. The ASTRO vehicle was equipped with a robotic arm to perform satellite capture and ORU exchange operations. After its launch in March 2007, various mission scenarios were successfully conducted, including visual

Fig. 19.4 Robonaut 2. *Image*
NASA



inspection, fuel transfer, ORU exchange, fly-around, rendezvous, docking and satellite capture. The free-flying capture was conducted autonomously using vision-based feedback [7].

19.1.1.5 Robonaut

Robonaut is a dexterous humanoid robot designed and built at NASA's Johnson Space Center in the United States. Building machines that can assist humans to work in and explore space is a key challenge. The Robonauts were designed to accomplish dexterous manipulation tasks using sophisticated human-like hands with tendon-driven fingers possessing multiple DOFs. The goal was to achieve dexterity that exceeds that of a suited astronaut. The advantage of a human-like robot is that the same workspace and tools designed for crewed space missions can be used. This not only improves efficiency, but also removes the need for specialized tools or interfaces for performing robotic operations.

Work on the first Robonaut began in 1997, and the first model called Robonaut 1 (R1), came out in 2002. Through 2006, R1 performed numerous experiments in a variety of laboratory and field test environments, proving that the concept of a robotic assistant was valid. The second generation Robonaut 2 (R2), was revealed in 2010, see Fig. 19.4. It is more technologically advanced than R1 and was delivered to the ISS by STS-133 in February 2011, becoming the first humanoid orbital robot on-board the ISS [8].

The Robonaut is a human-torso-like robot that contains joints with a total of 42 DOFs. Each arm has 7 DOFs, with a hand that has 12-DOF fingers. All the actuators are mounted in the arm. The torso contains 38 Power PC processors. There are more than 350 sensors in total, which are used for force/torque control based dexterous manipulation, as well as for safety behaviors.

Although, at present R2's primary role on the space station is limited to experiments inside the Destiny laboratory, the future enhancement plan includes the incorporation of a lower body to allow it to move around the station's interior. In addition, future upgrade could enable it to move outside to help astronauts with EVA tasks or perform repairs on the exterior of the station. Combined with a surface mobility system like legs or wheels, R2 could perform as a human-like manipulation system for future exploration missions on the Moon or Mars.

Orbital space robots will be able to assist humans in space by constructing and maintaining space modules and structures. Robotic manipulators have played essential roles in orbital operations. Moreover, satellite servicing missions are crucial to prevent the increase of space debris. The concept of servicing robots, or free-flying robots, has been discussed for many years, but there has been a limited number of validation flights in orbit, so far. More technological developments are expected to realize free-flying robots for servicing, rescuing or capture-and-removal missions of existing spacecraft in orbit.

19.1.2 Planetary Robots

19.1.2.1 Apollo 'Moon Buggy' and Lunokhod

The research on lunar surface mobility systems, which represents the roots of today's exploration rovers, began in the 1960s, with an initiative to develop a crewed roving vehicle ('Moon buggy', see Fig. 19.5) for the Apollo program in the United States, along with that for a teleoperated rover called Lunokhod in the Soviet Union. Both the Apollo rovers (Apollo 15–17 in 1971–1972) and the Lunokhod rovers (Lunokhod 1 in 1970 and Lunokhod 2 in 1973) were successfully operated on the Moon [9].

Fig. 19.5 Astronaut Eugene A. Cernan, mission commander, makes a short checkout of the lunar roving vehicle (LRV) during the early part of the first Apollo 17 Extravehicular Activity at the Taurus-Littrow landing site on December 11, 1972. Image NASA



There were numerous engineering design issues that had to be overcome to make vehicles work in this extraterrestrial environment, which contains high radiation, vacuum, severe temperatures and irregular terrain covered with regolith and dust. This was particularly true for the Lunokhod rovers, which had a mass of 840 kg with eight wheels supported by a dedicated suspension mechanism, and traveled 10.5 km (Lunokhod 1) and 37 km (Lunokhod 2) over the lunar terrain via television-image-based teleoperation from the ground station. To keep the rover warm during the long lunar nights, a polonium-210 radioactive heat source was successfully used.

19.1.2.2 Mars Landers: From Viking to Phoenix

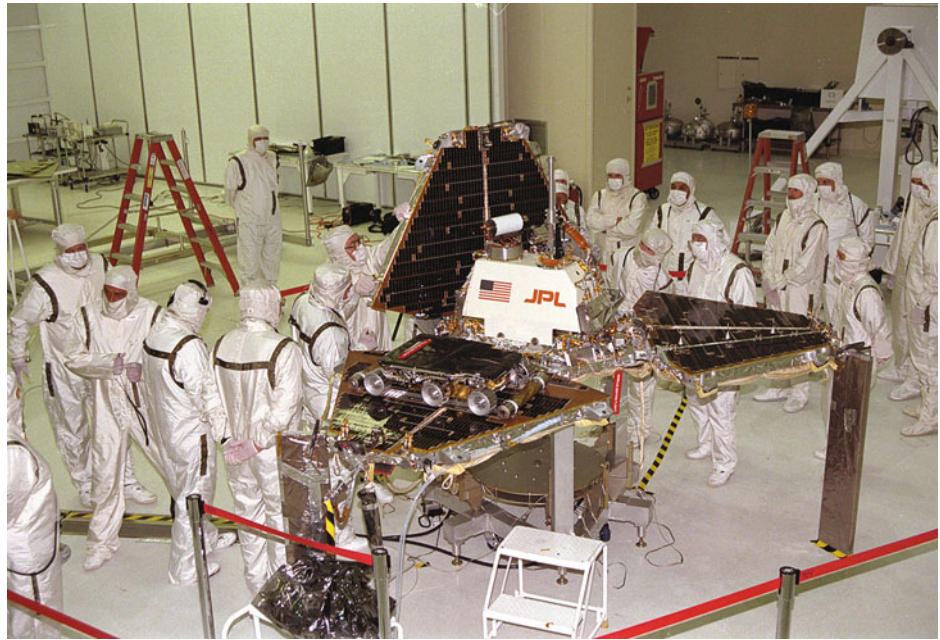
Upon the success of the lunar programs, the exploration target shifted to Mars. In 1976, two Viking landers (Viking 1 and Viking 2) developed by NASA landed on the surface of Mars. They each had a simple robotic arm to collect surface soil samples and put them into on-board containers for in situ analysis. After the Viking mission, there were multiple missions that were planned and actually launched to Mars, but it took about 30 years until the next successful lander mission. The Mars Phoenix Lander that successfully landed in a polar region of the Mars in 2008 had a much more sophisticated robotic arm. This robotic arm was operated to dig trenches in the Martian regolith and to acquire (scoop) dry and icy soil samples and deliver them to the in situ analyzers. It was also able to insert a sensor probe into the soil, and to position sensors and cameras at various locations near the lander.

Meanwhile, the Soviet Union also developed multiple missions to Mars, including orbiters, landers and rovers. In 1971, the Mars 2 and 3 missions successfully arrived in Martian orbit and attempted soft landings of both landing modules, which included a miniature rover; Mars 2 crashed on the surface, and Mars 3 lost communication soon after the landing. In 1988, two lander missions to Phobos, a moon (satellite) of Mars were launched in the Soviet Phobos program; Phobos-1 suffered a terminal failure *en route* to Mars, while Phobos-2 attained Mars orbit and returned 38 images of Phobos with a resolution of up to 40 m, but contact was lost prior to deployment of a planned Phobos lander. Later, Russia also developed the Mars-96 mission, which included an orbiter, lander and penetrator, but failed at launch. Along with that, a landing and rover mission was planned and the technology, including a rover testbed called Marskhod, was developed, but was not launched.

19.1.2.3 Mars Rovers: Pathfinder, MER and MSL

Autonomous or semi-autonomous robotic vehicles are considered as indispensable technology for planetary exploration. As a precursor mission for mobile robotics technology on a remote planet, the Mars Pathfinder mission deployed a micro-rover called *Sojourner* in 1997, see Fig. 19.6. The *Sojourner* rover traversed the rocky Martian surface in close vicinity to the landing site by autonomously avoiding obstacles [10]. Based on this successful technology demonstration, NASA developed larger, more capable twins for the Mars Exploration Rover (MER) mission, see Fig. 19.7, both of which were launched in 2003. The MER-A rover (*Spirit*)

Fig. 19.6 In spacecraft assembly and encapsulation facility-2 (SAEF-2), Jet Propulsion Laboratory workers are closing up the metal ‘petals’ of the Mars Pathfinder lander. The *Sojourner* small rover is visible on one of the three petals. Image NASA



landed on the Gusev crater on January 4, 2004, and the MER-B rover (*Opportunity*) landed on the Meridiani Planum on the opposite side of Mars from *Spirit* on January 25, 2004.

Both Pathfinder and the MER rovers introduced new technologies. Firstly, for the landing, a combination of an aerodynamic parachute and a unique airbag system was developed. Compared to a conventional lander, which uses a powered descent and soft landing, the airbag system can greatly reduce the mass of the landing module and its fuel, although it eliminates the precision landing feature by allowing the lander to bounce around on the surface several times before it finally settles down at a certain position.

Secondly, to achieve rough terrain mobility, these rovers use six independently driven wheels connected by a unique suspension arrangement called the rocker-bogie system. The term ‘rocker’ comes from the design of the differential that keeps the rover body balanced, enabling it to ‘rock’ depending on the various positions of the multiple wheels. The term ‘bogie’, on the other hand, comes from the old railroad systems and refers to a train undercarriage with six wheels that can swivel to curve along a track. To achieve this performance, the axles of the six wheels are connected by a passive linkage mechanism, with no need for springs, dampers, or even active elements. Thanks to this mechanism, the rover can move over a rock obstacle that is larger than the diameter of the wheel. The six-wheel and rocker-bogie suspension design was also adopted for NASA’s next rover (*Curiosity*) in the Mars Science Laboratory, which landed on Mars in 2012.

The MER rovers *Spirit* and *Opportunity* have an on-board manipulator arm for scientific operations. At the tip of this arm, several attached instruments can be placed directly up

against a rock or soil target of interest. For example, by using a rock abrasion tool, the surface of a rock can be scrubbed, after which the interior of the rock can be carefully observed using a microscopic camera and an alpha-particle X-Ray spectrometer. On-board the MER rovers, a stereo pair of high-resolution color CCD cameras are also mounted at the top of the Pancam Mast Assembly. This allows the cameras to rotate a full 360° to obtain a panoramic view of the Martian landscape. The stereoscopic measurement is used for mapping of the surrounding environment and as a vision-based odometry system for rover navigation [11].

The *Sojourner* rover weighs about 10.5 kg and is approximately the size of a microwave oven, the *Spirit* and *Opportunity* rovers weigh about 175 kg and are the size of golf carts, and the *Curiosity* rover weighs about 900 kg and is the size of a car. The *Sojourner* rover was actively operational for almost 3 months and traveled approximately 100 m in total. The mission of the *Spirit* rover was terminated in May 2011 after more than 7 years of operation on the surface. The total traveling distance was 7.73 km. On the other hand, the *Opportunity* rover remained operational throughout 2012 into 2013, with a cumulative distance traveled of more than 30 km [12].

The *Curiosity* rover landed on the Gale Crater on Mars on 6 August, 2012, see Fig. 19.8. As it is much heavier than *Sojourner*, *Spirit* and *Opportunity* and a much more precise landing was demanded, it used an innovative soft-landing system that combined parachute descent, powered descent and finally a ‘sky-crane’ to lower the rover to the surface on a tether. Despite its great complexity, the landing was successful at almost the center of the ellipsoid target area of about 6 km by 20 km [13].

Fig. 19.7 Artist's rendering of a Mars Exploration Rover. *Image* Maas Digital LLC for Cornell University and NASA/JPL

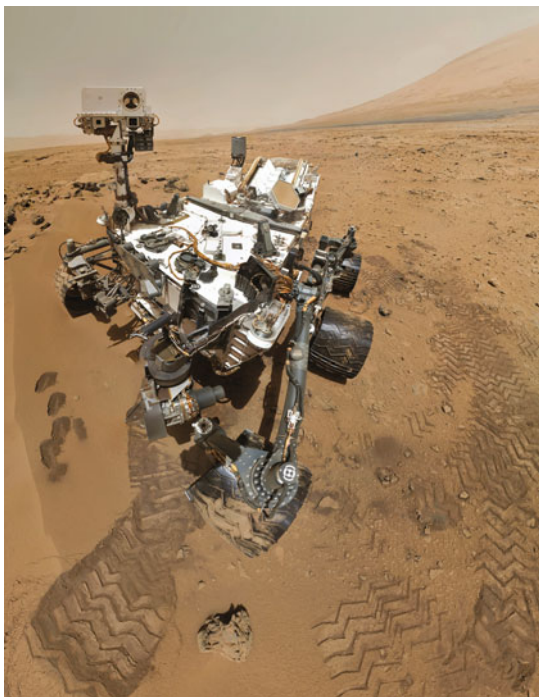


Fig. 19.8 A self-portrait by NASA's *Curiosity* rover in Gale Crater using the Mars Hand Lens Imager (*MAHLI*) to capture this set of 55 high-resolution images stitched together to create this full-color image on 31 October, 2012. *Image* NASA/JPL-Caltech/Malin Space Science Systems

19.1.2.4 Robotic Probes to Minor Celestial Bodies

In our solar system, there are numerous minor celestial bodies, such as asteroids, comets and satellites of the major planets, and the investigation of those bodies is also valuable for science. When comet Halley returned to the vicinity of the Sun (perihelion) in 1986, multiple space probes, including the European spacecraft *Giotto*, were launched to conduct detailed observations of the structure of the comet

nucleus and the mechanism of coma and tail formation. As for asteroids, the first successful mission to rendezvous and long-term observe one was NEAR-Shoemaker, which was launched in 1996 and arrived at the asteroid 433 Eros in 2000. Scientific observation continued until the craft finally touched down on the surface of Eros in 2001. Other minor body missions include Deep Space 1 (NASA, launched 1998), Stardust (NASA, launched 1999), Contour (NASA, launched 2002 but failed), Rosetta (ESA, launched 2004), Deep Impact (NASA, launched 2005), Dawn (NASA, launched 2007) and Hayabusa (ISAS/JAXA, launched 2003).

Hayabusa was to visit a near-Earth asteroid, acquire sample materials from its surface and return them to Earth for detailed analysis. It was developed by the Japanese Institute of Space and Astronautical Science (ISAS), which later became a part of JAXA. The probe was launched in May 2003 from Uchinoura Space Center, Japan and its re-entry capsule safely returned to the Woomera Desert of Australia in June 2010, successfully returning dust-like soil samples of the target asteroid 25143 Itokawa.

Sample-return is the method of bringing material back from space instead of taking analysis equipment all the way to space. It is the most difficult and ultimate probing method. However, the material can be analyzed with greater precision using the latest technology on Earth, even if the specimen is very small.

To achieve the *Hayabusa* sample-return mission, the following three innovative technologies were developed. The first was an ion engine (electric propulsion system). *Hayabusa* was equipped with four sets of newly developed cathode-less but microwave-discharge ion engines for the round trip mission to the target. A single engine had a nominal performance of 8 mN of thrust, with 3,000 s of specific impulse. The ion propulsion system worked

effectively throughout its 7 year deep space mission. The total accumulated operational time reached almost 40,000 h for all four ion engines, which consumed 47 kg of xenon propellant and provided a total ΔV of 2,200 m/s [14].

The second innovative technology was an autonomous optical navigation system for conducting a rendezvous maneuver with Itokawa, and then a touch-down operation on a specific location on the surface of this tiny object ($535 \times 294 \times 209$ m) located at a distance of 300,000,000 km from Earth, requiring approximately 33 min (2,000 s) for a round-trip communication [15].

The third technology involved material sampling in a microgravity field. The gravity field on Itokawa's surface is estimated to be about 100,000 times less than that of Earth's. This requires a far lower fuel consumption for performing landing and liftoff maneuvers compared to those performed on major planets or the Moon, but the lack of gravity makes it difficult to remain in place on the surface and acquire samples. Therefore, a 'touch-and-go' type of sample acquisition system was developed [16].

The Rosetta mission was launched in 2004 to the comet 67P/Churyumov-Gerasimenko, targeting an encounter in 2014. The mission objective is to travel to and land upon the surface of the comet to study its nucleus. The Rosetta probe is equipped with specially a designed anchor system and drilling mechanism to drill the comet's surface materials and conduct in situ analysis.

The robots that can land and travel on the lunar or planetary surfaces have been greatly contributing to our knowledge of the solar system. Wheeled mobile vehicles or robot rovers are successful on the natural and rough surface terrains of the Moon and Mars. Minor celestial bodies, such as asteroids and comets, have been also visited by many space probes. Minor bodies are characterized by very weak gravity fields; this fact makes the approach and landing maneuvers relatively easier, but the surface locomotion difficult.

19.2 Modeling and Control of Orbital Space Robots

Orbital robots are similar to terrestrial robots in that they are machines composed of multiple links jointed together to form arm-like structures, called manipulators, which are capable of performing a variety of tasks with specialized end-effectors and tools. The joints of the manipulator(s) are usually designed as single-degree-of-freedom (DOF) rotational joints driven by the appropriate actuators.

From the perspective of modeling and controlling orbital robots, it is appropriate to distinguish between extra-vehicular and intra-vehicular orbital robots. On the one hand, representative examples of extra-vehicular robots are

SRMS/SSRMS/JEMRMS and ETS-VII/Orbital Express; on the other hand, an example of an intra-vehicular robot is ROTEX. Extra-vehicular robots may pose more challenging modeling and control problems than intra-vehicular robots, because the latter resemble terrestrial robots to a higher degree. Indeed, large-workspace manipulators such as the SRMS on the Space Shuttle and the SSRMS/JEMRMS on the International Space Station are known to exhibit structural vibrations due to the specific design constraints imposed mainly on their mass [18]. Modeling a robot as flexible-link [19] and/or flexible-joint [20] manipulators and employing the respective methods of control is crucial for minimizing the vibrations [21–24].

Further, and as noted in Sect. 19.1, smaller manipulators can be attached to the end-links of the SSRMS and the JEMRMS (SPDM/Dextre and the Small Fine Arm, respectively), thus forming a 'macro-mini' manipulator structure. This leads to further challenges in terms of modeling and robot control. The motions of the mini-manipulator(s) may induce structural vibrations in the large arm, the joints of which remain locked during mini-manipulator operations. In this case, a *flexible-base manipulator* model would be appropriate. Hence, a controller must be designed that minimizes the reactions imposed on the flexible base from the mini-manipulator motions, and/or damps the excited vibrations (i.e. active damping via the mini-manipulator) [25].

Another class of extra-vehicular orbital robots are free-flying robots, e.g. the Space Shuttle with SRMS, ETS-VII or Orbital Express, that comprise a manipulator arm mounted on a satellite base. The base can attain any position and orientation depending on the forces and moments acting on it. The maneuvering capability of the satellite base can be achieved in the conventional way, i.e. using jet thrusters and the attitude control system (ACS). Similar to flexible-base robots, the acting forces and moments on the satellite base will also include undesirable reactions when set into motion by the manipulator arm. From the viewpoint of a conventional ACS, these forces are to be regarded as disturbances. However, such disturbances may not always be accommodated by the ACS, i.e. when there are inappropriate manipulator accelerations and/or large unknown payloads. One possibility to deal with this problem is to deactivate the ACS and let the base float freely during manipulator operation [26]. However, as was the case with the ETS-VII and Orbital Express, which used teleoperation from a remote site (Earth), precise orientation of the satellite base is required for communication. Hence, special controller design must be realized in order to minimize the manipulator reactions [27].

This section considers the modeling and control problems of free-floating space robots and 'macro-mini' structures modeled as flexible-base manipulators. Modeling

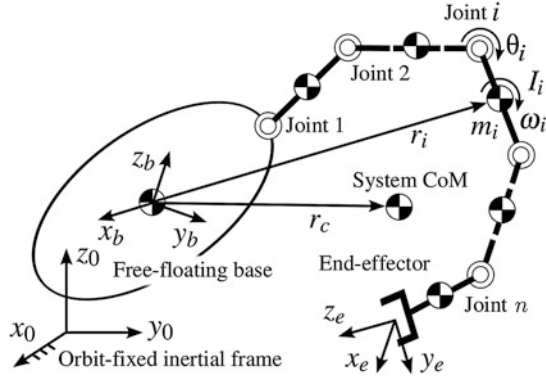


Fig. 19.9 Model of free-floating orbital space robot

issues are discussed in the first five subsections, including the underlying kinematic and dynamic equations, the system linear and angular momenta, two modeling approaches for free-floating robots in Sect. 19.2.3, the Reaction Null Space that is useful for disturbance minimization, and alternative dynamics formulations regarding ignorable coordinates, contact dynamics and extension to multi-arm robots, in Sect. 19.2.5. The last five subsections are devoted to basic control methods: end-link trajectory tracking control, point-to-point motion and non-holonomic path planning for free-floating robots, vibration suppression control for flexible-base robots, end-link impacts and impedance control, and post-impact control for momentum redistribution with regard to free-floating robots.

19.2.1 Kinematic and Dynamic Equations

Assume that the orbital robot is made of rigid-body links connected via n single-DOF joints. The joint coordinates will be denoted by $\theta \in \mathbb{R}^n$. The system can then be described with $6 + n$ generalized coordinates $q = (\mathcal{X}, \theta)$, where $\mathcal{X} \in SE(3)$ denotes the position/orientation of the satellite base w.r.t. an appropriately chosen inertial coordinate frame (usually assumed to be orbit-fixed).

First, the equation of motion for a free-flying space robot comprising a serial-link manipulator arm mounted on a satellite base is introduced (cf. Fig. 19.9). The equation is conveniently represented in the following block-matrix form

$$\begin{bmatrix} \mathbf{M}_b & \mathbf{M}_{bm} \\ \mathbf{M}_{bm}^T & \mathbf{M}_m \end{bmatrix} \begin{bmatrix} \dot{\mathcal{V}}_b \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_b \\ \mathbf{c}_m \end{bmatrix} = \begin{bmatrix} \mathcal{F}_b \\ \tau \end{bmatrix} + \begin{bmatrix} {}^b\mathbf{T}_e^T \\ \mathbf{J}_m^T \end{bmatrix} \mathcal{F}_e \quad (19.1)$$

where

\mathbf{M}_m	$\in \mathbb{R}^{n \times n}$: fixed-base manipulator link inertia matrix
\mathbf{M}_b	$\in \mathbb{R}^{6 \times 6}$: system articulated body inertia matrix
\mathbf{M}_{bm}	$\in \mathbb{R}^{6 \times n}$: coupling inertia matrix
\mathbf{c}_m	$\in \mathbb{R}^n$: fixed-base manipulator link Coriolis and centrifugal forces
\mathbf{C}_b	$\in \mathbb{R}^6$: Coriolis and centrifugal forces on the system articulated body
τ	$\in \mathbb{R}^n$: manipulator joint torque vector
\mathcal{V}_b	$\in \mathbb{R}^6$: spatial velocity of the base
$\mathcal{F}_b, \mathcal{F}_e$	$\in \mathbb{R}^6$: spatial forces on the base and the end-link, respectively
${}^b\mathbf{T}_e$	$\in \mathbb{R}^{6 \times 6}$: spatial coordinate transform
\mathbf{J}_m	$\in \mathbb{R}^{6 \times n}$: fixed-base manipulator Jacobian matrix

The lower-case bold characters denote vectors; the upper-case bold characters represent matrices; and the spatial quantities such as the rigid body spatial velocity and spatial forces are denoted by calligraphic symbols, e.g. $\mathcal{V}_O, \mathcal{F}_O \in \mathbb{R}^6$, respectively. The convention for spatial vectors composed of 3D quantities is as follows: a linear component followed by an angular component, e.g. $\mathcal{V}_O = [\mathbf{v}_O^T \ \omega^T]^T$ and $\mathcal{F}_O = [\mathbf{f}^T \ \mathbf{n}_O^T]^T$ where $\mathbf{v}, \omega, \mathbf{f}, \mathbf{n}$ denote 3D vectors of body velocity, angular velocity, force and moment, respectively. Spatial transforms are represented as

$${}^k\mathbf{T}_l = \begin{bmatrix} {}^k\mathbf{R}_l & -{}^k\mathbf{R}_l {}^k\mathbf{R}_l^\times \\ 0 & {}^k\mathbf{R}_l \end{bmatrix} \quad (19.2)$$

with ${}^k\mathbf{R}_l \in \mathbb{R}^{3 \times 3}$ denoting the orientation of coordinate frame $\{l\}$ with respect to $\{k\}$ and ${}^k\mathbf{R}_l^\times \in \mathbb{R}^{3 \times 3}$ denoting the skew-symmetric operator associated with the vector ${}^k\mathbf{r}_l \in \mathbb{R}^3$ that expresses the position of $\{l\}$ with respect to $\{k\}$.

The upper part of the above equation denotes the system articulated-body dynamics. The coordinates are those of the satellite base, but the inertial properties are those of the entire system, hence the term ‘articulated body’ [28]. The lower part of the above equation describes the dynamics of the manipulator. Because base coordinates were used, the quantities \mathbf{M}_m , \mathbf{c}_m and \mathbf{J}_m are those of the respective fixed-base manipulator. Furthermore, the entire equation includes components for the intercoupled inertial and nonlinear generalized forces on the left-hand side, and the external and/or driving forces on the right-hand side.

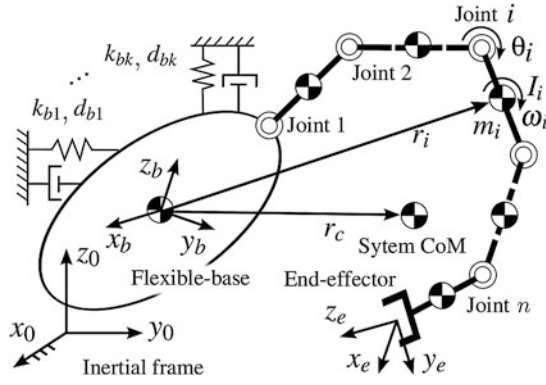


Fig. 19.10 Model of flexible-base manipulator system

For the case of a flexible-base space robot (cf. Fig. 19.10), two additional terms are added on the left-hand side to account for the base spatial damping and stiffness: they are expressed via diagonal matrices $\mathbf{D}_b, \mathbf{K}_b \in \mathbb{R}^{6 \times 6}$ with elements d_{bk} and $k_{bk}, k = 1, 2, \dots, 6$, respectively

$$\begin{bmatrix} M_b & M_{bm} \\ M_{bm}^T & M_m \end{bmatrix} \begin{bmatrix} \dot{\mathcal{V}}_b \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} C_b \\ c_m \end{bmatrix} + \begin{bmatrix} \mathbf{D}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathcal{V}_b \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta \mathcal{X}_b \\ \Delta \theta \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \tau \end{bmatrix} + \begin{bmatrix} {}^b \mathbf{T}_e^T \\ \mathbf{J}_m^T \end{bmatrix} \mathcal{F}_e. \quad (19.3)$$

The base external/driving force \mathcal{F}_b was set to zero. The kinematic equation for the velocity is given as

$$\mathcal{V}_e = T_{eb} \mathcal{V}_b + J_m(\theta) \dot{\theta} \quad (19.4)$$

where \mathcal{V}_e is the spatial velocity of the end-link. The first component on the right-hand side represents the base motion, and the second component represents the manipulator motion with respect to the base.

19.2.2 Linear and Angular Momenta

The spatial momentum of a free-floating robot consists of two elements: a linear and an angular one. The angular momentum component is written with respect to the center of mass (CoM) of the articulated body

$$\mathcal{L}_c \equiv \begin{bmatrix} \mathbf{p} \\ \mathbf{l}_c \end{bmatrix} = \mathbf{M}_c \mathcal{V}_c. \quad (19.5)$$

The linear element is $\mathbf{p} = \sum_{i=0}^n m_i \dot{\mathbf{r}}_i = m_t \dot{\mathbf{r}}_c$ and the angular element is $\mathbf{l}_c = \sum_{i=0}^n (\mathbf{I}_i \omega_i + m_i \mathbf{r}_i \times \dot{\mathbf{r}}_i)$ where $\mathbf{I}_i, m_i, \mathbf{r}_i, \omega_i$, represent the link i inertia matrix, mass, CoM position and angular velocity, respectively: all of which are in inertial coordinates. In addition, m_t denotes the mass of the articulated body system, and \mathbf{r}_c and \mathcal{V}_c denote its CoM

position and spatial velocity, respectively. The matrix \mathbf{M}_c is a block-diagonal matrix including $m_t \mathbf{U}$ and $\mathbf{I}_c \equiv \sum_{i=0}^n (\mathbf{I}_i - m_i \mathbf{R}_{ci}^x \mathbf{R}_{ci}^x)$ as upper and lower blocks, respectively.

Redefining spatial momentum with respect to the base gives

$$\mathcal{L}_b = \begin{bmatrix} \mathbf{p} \\ \mathbf{r}_{bc} \times \mathbf{p} + \mathbf{l}_c \end{bmatrix} \quad (19.6)$$

where \mathbf{r}_{bc} denotes the position of the articulated body CoM with respect to the base frame and yields the advantage of the application of familiar fixed-base manipulator inertial properties. This representation can be related to the equation of motion Eq. 19.1, as follows. Extracting the section from Eq. 19.1 that concerns the system articulated-body dynamics yields the following

$$\mathbf{M}_b \dot{\mathcal{V}}_b + \mathbf{M}_{bm} \ddot{\theta} + C_b = \mathcal{F}_{qs} \quad (19.7)$$

where $\mathcal{F}_{qs} = \mathcal{F}_b + \mathbf{T}_{eb}^T \mathcal{F}_e$ denotes the *quasistatic forces*. The dynamic equilibrium of the articulated-body system can be then expressed as $\mathcal{F}_d - \mathcal{F}_{qs} = \mathbf{0}$. Then the *dynamic force* \mathcal{F}_d can be obtained as the time derivative $\mathcal{F}_d = \frac{d}{dt} \mathcal{L}_b$. In the absence of quasistatic forces, i.e. when $\mathcal{F}_{qs} = \mathbf{0}$ and when the base is unactuated and no external forces act on the end-link, the articulated-body dynamics Eq. 19.7 can be integrated

$$\mathbf{M}_b \mathcal{V}_b + \mathbf{M}_{bm} \dot{\theta} = \bar{\mathcal{L}}_b \quad (19.8)$$

where $\bar{\mathcal{L}}_b$ is the integration constant. The first component on the left-hand side, $\mathbf{M}_b \mathcal{V}_b$, is the articulated-body momentum due to the base motion. The second component, $\mathbf{M}_{bm} \dot{\theta}$, is due to the manipulator motion. It plays an important role in path planning and control as will be shown below. The component is called *coupling momentum* [29] and will be denoted as \mathcal{L}_{bm} . It gives rise to a spatial force imposed on the base via manipulator motion

$$\mathcal{F}_{bm} \equiv \mathbf{M}_{bm} \ddot{\theta} + \dot{\mathbf{M}}_{bm} \dot{\theta}. \quad (19.9)$$

\mathcal{F}_{bm} will be henceforth referred to as the *imposed force*. Then, the articulated-body dynamics of a free-flying space robot in a form familiar from Newtonian mechanics can be represented as

$$\mathbf{M}_b \dot{\mathcal{V}}_b = -\mathcal{F}_{bm} \quad (19.10)$$

which was obtained from Eq. 19.1 under the assumption of no external forces, and the approximation of $C_b \approx \dot{\mathbf{M}}_{bm} \dot{\theta}$ [29].

Looking further for integrability of the momentum equation, the linear part is integrable, whereas the angular part is not. Hence, the latter represents a *non-holonomic constraint*, implying the orientation of the base cannot be

expressed as a function of the current manipulator joint angles; rather, it will depend on the history of the joint angle vector.

The articulated-body dynamics of a flexible-base robot have the same form as in Eq. 19.7, with the addition of quasistatic forces

$$\mathcal{F}_{qs} = \mathbf{T}_{eb}^T \mathcal{F}_e - \mathbf{D}_b \mathcal{V}_b - \mathbf{K}_b \Delta \mathcal{X}_b. \quad (19.11)$$

Even with no external force ($\mathcal{F}_e = 0$), the quasistatic forces will be non-zero, e.g. when the base is displaced from the equilibrium position because of the manipulator reaction. Hence, momentum conservation does not necessarily hold in this case. The articulated-body dynamics can be rewritten in the classical mass-damper-spring form via the above imposed force notation

$$\mathbf{M}_b \dot{\mathcal{V}}_b + \mathbf{D}_b \mathcal{V}_b + \mathbf{K}_b \Delta \mathcal{X}_b = -\mathcal{F}_{bm}. \quad (19.12)$$

19.2.3 Virtual Manipulator and Generalized Jacobian

A free-flying robot with an unactuated base obeys the law of momentum conservation. This is a special case: the dynamics are simplified, and additionally, velocity-based relations play a predominant role. However, inertial properties are involved in these relations, which is in contrast with the case of fixed-base terrestrial robots. Because the base is unactuated, it moves in reaction to manipulator motions. This results in a diminishment of the motion ability of the end-link and the workspace of the manipulator when compared to the same manipulator mounted on a fixed base.

There are two convenient concepts for dealing with such velocity-level models: the Virtual Manipulator [30] and the Generalized Jacobian [31]. The Virtual Manipulator has a massless kinematic chain fixed at the ‘virtual ground’—a point that does not move (under zero initial momentum) in inertial space. This point is the CoM of the articulated body system. Furthermore, the link lengths of the Virtual Manipulator depend on the inertial properties, if the joint arrangement matches that of the real manipulator, and if the joint axes are parallel to the respective axes of the real manipulator. With this construction, the degraded end-link motion ability due to the base motion can be accounted for.

Another convenient notation for velocity-level relations is the Generalized Jacobian. Spatial momentum conservation, as in Eq. 19.8, can be used as a *constraint* with respect to the manipulator motion. From Eq. 19.8, the base velocity is obtained as

$$\mathcal{V}_b = \bar{\mathcal{V}}_b - \mathbf{M}_b^{-1} \mathbf{M}_{bm} \dot{\theta} \quad (19.13)$$

where $\bar{\mathcal{V}}_b = \mathbf{M}_b^{-1} \bar{\mathcal{L}}_b$ is acquired from the initial spatial momentum and the second component is attributed to the coupling momentum induced by the manipulator motion. Inserting \mathcal{V}_b into Eq. 19.4, the constrained manipulator end-effector velocity is

$$\mathcal{V}_e = \bar{\mathcal{V}}_e + \hat{\mathbf{J}} \dot{\theta} \quad (19.14)$$

where $\bar{\mathcal{V}}_e = \mathbf{T}_{eb} \bar{\mathcal{V}}_b$. The matrix

$$\hat{\mathbf{J}} \equiv \mathbf{J}_m - \mathbf{T}_{eb} \mathbf{M}_b^{-1} \mathbf{M}_{bm}$$

is called the *Generalized Jacobian*.

19.2.4 The Reaction Null Space

As shown previously, the motion of the base in reaction to manipulator motion diminishes the end-link motion ability and the effective workspace. One possibility to mitigate this is to use custom path planning and control methods for manipulator motions that would minimize the reaction at the base. In fact, it is straightforward to predict the existence of *reactionless motion*. In other words, there are manipulator motions that will guarantee full dynamical decoupling between the base and the manipulator. This condition is expressed simply as $\mathcal{F}_{bm} = \mathbf{0}$.

When the system articulated-body dynamics Eq. 19.7 of an orbital space robot with an unactuated base, zero initial base velocity ($\bar{\mathcal{V}}_b = \mathbf{0}$), and zero external forces ($\mathcal{F}_{qs} = \mathbf{0}$) is considered with Eqs. 19.10 or 19.12, the following relation results

$$\mathcal{F}_{bm} = \mathbf{M}_{bm} \ddot{\theta} + \dot{\mathbf{M}}_{bm} \dot{\theta} = \mathbf{0} \quad (19.15)$$

where the nonlinear force C_b in (19.7) was approximated as it was in Eq. 19.10. This equation can be integrated once to obtain the momentum equation

$$\mathbf{M}_{bm} \dot{\theta} = \bar{\mathcal{L}}_{bm} \quad (19.16)$$

where \mathcal{L}_{bm} denotes the coupling momentum. This is a linear equation for the velocities and its solution type depends on the number of manipulator joints n . The equation will be determined if $n = 6$, and under-determined otherwise ($n > 6$). In the latter case, the joint velocity vector derived from the above equation is

$$\dot{\theta} = \mathbf{M}_{bm}^+ \bar{\mathcal{L}}_{bm} + \mathbf{P}_{M_{bm}} \dot{\theta}_a \quad (19.17)$$

where $(\circ)^+$ is the Moore–Penrose generalized inverse, $\mathbf{P}_{(\circ)}$ is a null-space projector and $(\circ)_a$ is an arbitrary vector [29]. The two components on the r.h.s. are orthogonal, implying that any joint velocity from the null space of the coupling inertia matrix will not change the momentum of the *base*.

These types of manipulator motions are termed *reactionless* and are obtained by varying the arbitrary velocity vector $\dot{\theta}_a$. The null space itself is termed the *Reaction Null Space* (RNS) [29] and is useful for motion analysis, path planning and reactionless motion control.

The set of reactionless motions depends on the rank of the RNS projector: $\text{rank} \mathbf{P}_{M_{bm}} = n - 6$. With a seven-DOF manipulator, e.g., the set will be just one-dimensional, implying that reactionless motions are possible only along the integral curves of the above differential equation. In general, it is desirable to have a larger set of such paths. One possibility to achieve this is to increase the number of manipulator joints (i.e. the DOFs). Another option is to redefine the RNS with respect to some of the base coordinates. From a practical viewpoint, the orientation of the base is the most important factor, hence, the RNS can be redefined only with respect to the angular variables. For that case, the rank of the RNS projector will increase to $n - 3$. An example is shown in Sect. 19.2.10.

19.2.5 Other Representations of System Dynamics

19.2.5.1 Ignorable Coordinates

From analytical mechanics it is known that conserved quantities in the equation of motion yield *ignorable or cyclic coordinates*. In the case of free-floating robot dynamics, such are the coordinates of the base. This property was already used when deriving the Generalized Jacobian in Eq. 19.14 from the kinematic and momentum equations. The ignorable coordinates can also be removed in a similar way from the dynamic equation Eq. 19.1. This leads to a representation in a reduced form

$$\hat{\mathbf{M}}_m \ddot{\theta} + \hat{\mathbf{c}} = \hat{\boldsymbol{\tau}} + \hat{\mathbf{J}}^T \mathcal{F}_e \quad (19.18)$$

where $\hat{\mathbf{M}}_m = \mathbf{M}_m - \mathbf{M}_{bm}^T \mathbf{M}_b^{-1} \mathbf{M}_{bm}$, $\hat{\mathbf{c}}_m = \mathbf{c}_m - \mathbf{M}_{bm}^T \mathbf{M}_b^{-1} \mathbf{C}_b$ and $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau} - \mathbf{M}_{bm}^T \mathbf{M}_b^{-1} \mathcal{F}_b$ [32]. The dimension of the equation is decreased to n and is the same as for a fixed-base manipulator.

Furthermore, system dynamics can be represented in terms of *quasi-coordinates* by using the articulated-body quasi-coordinates \mathcal{V}_c instead of the base coordinates \mathcal{X}_b . The articulated-body dynamics are derived via time differentiation of spatial momentum in Eq. 19.5

$$\mathbf{M}_c \dot{\mathcal{V}}_c + \mathbf{C}_c = \mathbf{T}_{ec}^T \mathcal{F}_e \quad (19.19)$$

where \mathbf{C}_c denotes the non-linear forces. Combing with the reduced dynamics in Eq. 19.18, the total dynamics in a decoupled form is as follows

$$\begin{bmatrix} \mathbf{M}_c & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{M}}_m \end{bmatrix} \begin{bmatrix} \dot{\mathcal{V}}_c \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_c \\ \hat{\mathbf{c}}_m \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \hat{\boldsymbol{\tau}} \end{bmatrix} + \begin{bmatrix} \mathbf{T}_{ec}^T \\ \hat{\mathbf{J}}^T \end{bmatrix} \mathcal{F}_e. \quad (19.20)$$

19.2.5.2 End-Link Contact Dynamics

During manipulation and other tasks, the end-link may establish contact with an object. Spatial contact forces will thereby be generated and subsequently propagated via the end-link to the rest of the robot links. Hence, it is crucial to model the end-link contact dynamics.

The spatial velocity of the end-link is represented in Sect. 19.4, which uses base coordinates as an intermittent frame. Because these are ignorable coordinates, the relation can be rewritten via the articulated-body quasi-coordinates as

$$\mathcal{V}_e = \mathbf{T}_{ec} \mathcal{V}_c + \hat{\mathbf{J}} \dot{\theta}. \quad (19.21)$$

To obtain the dynamic relations, the respective acceleration will be used

$$\dot{\mathcal{V}}_e = \mathbf{T}_{ec} \dot{\mathcal{V}}_c + \hat{\mathbf{J}} \ddot{\theta} + \dot{\mathbf{T}}_{ec} \mathcal{V}_c + \dot{\hat{\mathbf{J}}} \dot{\theta}. \quad (19.22)$$

The quasi-coordinate acceleration $\dot{\mathcal{V}}_c$ and the joint acceleration $\ddot{\theta}$ can be obtained from the articulated-body dynamics in Eq. 19.19 and from the reduced form of dynamics in Eq. 19.18, respectively. In contact scenarios, two cases are usually considered: free manipulator joints ($\boldsymbol{\tau} = \mathbf{0}$) and locked manipulator joints ($\dot{\theta} = \mathbf{0}$) [33]. The end-link contact dynamics can then be represented as

$$\dot{\mathcal{V}}_e = \mathbf{M}_*^{-1} \mathcal{F}_e + \mathcal{A}_* \quad (19.23)$$

where \mathcal{A}_* denotes non-linear velocity-dependent end-link acceleration and

$$\mathbf{M}_*^{-1} = \mathbf{T}_{ec} \mathbf{M}_c^{-1} \mathbf{T}_{ec}^T + \kappa \hat{\mathbf{J}} \hat{\mathbf{M}}_m^{-1} \hat{\mathbf{J}}^T \quad (19.24)$$

represents the *mobility tensor* s.t. $\kappa = 1$ in the free-joint case, and $\kappa = 0$ in the locked-joint case.

19.2.5.3 Extension to Multi-Arm Orbital Robots

When a free-flying space robot has l manipulator arms mounted on a base, the manipulators comprise a tree-like structure. Each manipulator arm has n_k joints, $k = 1, 2, \dots, l$, resulting in the total number of joints of $n = \sum_{k=1}^l n_k$. External forces may act on the base as well as on one or more of the end-links. The dynamic equation Eq. 19.1 then becomes

$$\begin{bmatrix} \mathbf{M}_b & \mathbf{M}_{bm} \\ \mathbf{M}_{bm}^T & \mathbf{M}_m \end{bmatrix} \begin{bmatrix} \dot{\mathcal{V}}_b \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_b \\ \mathbf{c}_m \end{bmatrix} = \begin{bmatrix} \mathcal{F}_b \\ \boldsymbol{\tau} \end{bmatrix} + \begin{bmatrix} \mathbf{T}_{eb}^T \\ \mathbf{J}_m^T \end{bmatrix} \mathcal{F}_e \quad (19.25)$$

where $\theta = [\theta_1^T \theta_2^T \dots \theta_l^T]^T$, $\tau = [\tau_1^T \tau_2^T \dots \tau_l^T]^T \in \mathbb{R}^n$, $\mathcal{F}_e = [\mathcal{F}_{e_1}^T \mathcal{F}_{e_2}^T \dots \mathcal{F}_{e_l}^T]^T \in \mathbb{R}^{6l}$, the Jacobian $\mathbf{J}_m \in \mathbb{R}^{6l \times n}$ is a block-diagonal with blocks $\mathbf{J}_{m_k} \in \mathbb{R}^{6 \times n_k}$ and is the fixed-base manipulator Jacobian of the k -th arm, \mathcal{F}_{e_k} is the spatial force acting at its end-link, and $\mathbf{T}_{eb}^T \in \mathbb{R}^{6 \times 6l}$ is composed of matrices $\mathbf{T}_{e_k b}^T$ [34–37].

The kinematic equations, on the other hand, can be written as

$$\mathcal{V}_{e_k} = \bar{\mathcal{V}}_{e_k} + \hat{\mathbf{J}}_k \dot{\theta}_k, \quad k = 1, 2, \dots, l \quad (19.26)$$

where \mathcal{V}_{e_k} is the spatial velocity of the k -th end-link, $\bar{\mathcal{V}}_{e_k} = T_{e_k b} \bar{\mathcal{V}}_b$ is a result of the initial spatial momentum and the matrices $\hat{\mathbf{J}}_k \equiv \hat{\mathbf{J}}_{m_k} - \mathbf{T}_{e_k b} \mathbf{M}_b^{-1} \mathbf{M}_{b m_k}$ are the Generalized Jacobians [34].

The dynamic equation Eq. 19.3 can be recast in a similar fashion when the flexible-base space robot includes more than one manipulator.

19.2.6 Velocity-Based End-Link Trajectory Tracking Control

Velocity-based control is used in teleoperation mode, as explained in Sect. 19.1. However, velocity-based end-link trajectory tracking is used in an autonomous mode of operation to accomplish precise motion tasks such as approaching specific parts of hardware equipment. The end-link path is planned in order, for example, to avoid undesirable interference with other parts of the equipment. Typically, feedback control would be employed in workspace coordinates based on the manipulator's inverse Jacobian [38]. Orbital robots can be directly controlled with such methods when the end-link trajectory is designed with respect to the base coordinate frame. For trajectories specified in inertial (orbit-fixed) coordinates (e.g. during satellite capturing, satellite repair or payload transfer), the base deflection due to reactions should be taken into account. For the case of an unactuated base, the feedback controller can be designed using the Generalized Jacobian formulation from the previous section [31]. The manipulator joint velocities to be used as control inputs for the velocity-level feedback controller are

$$\dot{\theta} = \hat{\mathbf{J}}^{-1} (\mathbf{K}_p (\mathcal{X}_e^d - \mathcal{X}_e) + \mathcal{V}_e^d) \quad (19.27)$$

where \mathcal{X}_e^d and \mathcal{V}_e^d denote the desired end-link spatial position and velocity along the given inertial trajectory and \mathbf{K}_p is a feedback gain matrix. The actual end-link position \mathcal{X}_e is obtained by summing up two components: the inertial base position, obtained via appropriate measurements, and the end-link position w.r.t. the base, obtained via the direct kinematics relations for fixed-base robots based on manipulator joint position measurements [38].

19.2.7 Point-to-Point Motion and Nonholonomic Path Planning

Point-to-point (PTP) motion control is a method of manipulator motion control that ensures precise positioning of the end-link at a desired spatial position in inertial space or attaining a desired manipulator configuration. In this case, the motion trajectory is of little interest [38]. The folding and unfolding of the manipulator arm to/from the stowed position is usually carried out via PTP motion control in joint space coordinates. Alternatively, tasks that require end-link positioning with respect to some equipment can be done either in base coordinates, in which the equipment is fixed to the base, or in inertial coordinates, in which the equipment is fixed to another body. In the latter case, PTP motion control can be realized via the Generalized Jacobian feedback control equation

$$\dot{\theta} = \hat{\mathbf{J}}^{-1} (\mathbf{K}_p (\mathcal{X}_e^d - \mathcal{X}_e) - \mathbf{K}_d \dot{\mathcal{X}}_e). \quad (19.28)$$

The base may thereby freely change its state.

Especially, in the case of a free-flying robot with an unactuated base, the system exhibits non-holonomic behavior owing to the nonintegrability condition on the spacecraft attitude. Nevertheless, it is possible to control the base attitude during PTP operations, e.g. via a bidirectional path planning method [39, 40].

19.2.8 Vibration Suppression Control

For flexible-base space robots, the vibrations of the base may lead to end-link task performance deterioration. It is possible to suppress the vibrations of the base via manipulator motion, using the inertial coupling between the base and the manipulator [25]. This becomes apparent when analyzing the articulated-body dynamics expressed from Eqs. 19.7 and 19.11 as

$$\mathbf{M}_b \dot{\mathcal{V}}_b + \mathbf{M}_{bm} \ddot{\theta} + C_b = \mathbf{T}_{eb}^T \mathcal{F}_e - \mathbf{D}_b \mathcal{V}_b - \mathbf{K}_b \Delta \mathcal{X}_b. \quad (19.29)$$

Additional damping can be injected into the above dynamics via a control joint acceleration [29]

$$\ddot{\theta} = \mathbf{M}_{bm}^+ (\mathbf{D}_{bc} \mathcal{V}_b - C_b) \quad (19.30)$$

where \mathbf{D}_{bc} denotes a matrix for additional damping. This confirms that in the absence of external forces ($\mathcal{F}_e = \mathbf{0}$), the following closed-loop dynamics are obtained

$$\mathbf{M}_b \dot{\mathcal{V}}_b + (\mathbf{D}_{bc} + \mathbf{D}_b) \mathcal{V}_b + \mathbf{K}_b \Delta \mathcal{X}_b = 0. \quad (19.31)$$

19.2.9 End-Link Impacts and Impedance Control

A task of utmost importance for orbital space robots is the retrieval of floating bodies, e.g. malfunctioned satellites or space debris. Because it is usually assumed that the target object lacks any dedicated grapple fixture, special care is needed when establishing the initial contact and selecting the post-contact tracking control method for the robot arm so that the target is not pushed away during the operation.

The inertial properties during the initial impact depend on the end-link contact dynamics, as described in Sect. 19.2.5. The end-link approach direction specified via a unit vector \mathbf{n} is assumed to be known. Therefore, the inertial properties can be described in terms of a scalar: the effective mass m_* with an impact along \mathbf{n} . This mass can be obtained from the mobility tensor \mathbf{M}_*^{-1} in Eq. 19.24 as follows

$$m_* = \frac{\|\mathbf{f}_e\|}{\mathbf{n}^T \dot{\mathbf{v}}_e} = \frac{1}{\mathbf{n}^T \mathbf{M}_{ff}^{-1} \mathbf{n}} \quad (19.32)$$

where \mathbf{f}_e and $\dot{\mathbf{v}}_e$ denote the linear parts of the spatial end-link force and acceleration, respectively, and \mathbf{M}_{ff}^{-1} is the upper-left 3×3 block sub-matrix of the mobility tensor. Because the tensor is manipulator configuration dependent for a given approach direction \mathbf{n} , the effective mass can be varied by changing the configuration at impact.

However, the effective mass variation via manipulator configuration is limited [41]. A broader range can be achieved with the help of *mechanical-impedance control*: a method suggested in [42] for fixed-base manipulator end-link control during contact tasks. The end-link dynamics are specified thereby via the equation

$$\mathbf{M}_e \dot{\mathcal{V}}_e + \mathbf{D}_e \mathcal{V}_e + \mathbf{K}_e \Delta \mathcal{X}_e = \mathcal{F}_e \quad (19.33)$$

where \mathbf{M}_e , \mathbf{D}_e and \mathbf{K}_e are desired mechanical-impedance related spatial transforms for inertia, damping and stiffness, respectively. These quantities determine the end-link behavior during contact. To ensure the above end-link dynamics, the following joint control torque is applied

$$\hat{\boldsymbol{\tau}} = (\hat{\mathbf{M}}_m \hat{\mathbf{J}}^{-1} \mathbf{M}_e^{-1} - \hat{\mathbf{J}}^T) \mathcal{F}_e + \hat{\mathbf{c}} - \hat{\mathbf{M}}_m \hat{\mathbf{J}}^{-1} \left[\mathbf{M}_e^{-1} (\mathbf{D}_e \mathcal{V}_e + \mathbf{K}_e \Delta \mathcal{X}_e) + \hat{\mathbf{J}} \dot{\boldsymbol{\theta}} \right]. \quad (19.34)$$

This equation was obtained using the reduced form of the dynamics in Eq. 19.18 and the kinematic relation in Eq. 19.21. Unfortunately, the equation is quite complex. Additionally, a high control bandwidth would be required to realize the desired end-link behavior [41].

A formulation for impedance control of multi-arm free-floating robots can be found in [43].

19.2.10 Post-Impact Control for Momentum Redistribution

The momentum transferred to the articulated body after the impact with the target may lead to a significant base translation or rotation. Rotation can be especially harmful and is highly undesirable. It is possible to employ the manipulator arm to accommodate a portion of the momentum transferred to the space robot via the impact, thus minimizing the initial post-impact base momentum [44]. The accommodated momentum can then be transferred to the base and mitigated thereafter with the assistance of a reaction or momentum wheel control subsystem. This requires proper post-impact momentum redistribution control: the underlying equations are derived as follows. Focusing on the base rotation, the system dynamics are to be rewritten using only base angular velocity quasi-coordinates. First, the translational coordinates of the base are eliminated from the momentum equation. The angular momentum with respect to the base's CoM can be written as

$$\mathbf{I}_b = \tilde{\mathbf{M}}_\omega \boldsymbol{\omega} + \tilde{\mathbf{M}}_{\omega m} \dot{\boldsymbol{\theta}} + \tilde{\mathbf{M}}_{\omega \phi} \dot{\boldsymbol{\phi}} \quad (19.35)$$

where $\tilde{\mathbf{M}}_\omega = \mathbf{M}_\omega + m_t \mathbf{R}_{bc}^\times \mathbf{R}_{bc}^\times$ and $\tilde{\mathbf{M}}_{\omega m} = \mathbf{M}_{\omega m} + \mathbf{R}_{bc}^\times \mathbf{M}_{vm}$. These block matrices are derived from the articulated-body system and the coupling inertia matrices $\mathbf{M}_b = \begin{bmatrix} \mathbf{M}_v & \mathbf{M}_{v\omega} \\ \mathbf{M}_{v\omega}^T & \mathbf{M}_\omega \end{bmatrix}$ and $\mathbf{M}_{bm} = [\mathbf{M}_{vm}^T \quad \mathbf{M}_{\omega m}^T]^T$, respectively. Detailed expressions for the sub-matrices can be found in [32]. $\tilde{\mathbf{M}}_{\omega \phi} \dot{\boldsymbol{\phi}}$ represents the angular momentum component due to the momentum wheels, and $\dot{\boldsymbol{\phi}}$ denotes the respective quasi-coordinates. The tilde operator modifies the respective matrix in such a way that linear motion of the base is implicitly accounted for.

Angular momentum is conserved during the post-impact phase. Hence, the manipulator control joint rates can be derived as

$$\dot{\boldsymbol{\theta}} = \tilde{\mathbf{M}}_{\omega m}^+ \left(\bar{\mathbf{I}}_b - \tilde{\mathbf{M}}_\omega \boldsymbol{\omega}^d - \tilde{\mathbf{M}}_{\omega \phi} \dot{\boldsymbol{\phi}} \right) + \mathbf{P}_{\tilde{\mathbf{M}}_{\omega m}} \dot{\boldsymbol{\theta}}_a^d \quad (19.36)$$

where $\bar{\mathbf{I}}_b$ denotes the conserved angular momentum. The articulated-body momentum component $\tilde{\mathbf{M}}_\omega \boldsymbol{\omega}^d$ and the RNS component $\mathbf{P}_{\tilde{\mathbf{M}}_{\omega m}} \dot{\boldsymbol{\theta}}_a^d$ can be used to minimize the base rotation and the joint motion, respectively, using damping controls $\boldsymbol{\omega}^d = -\mathbf{K}_\omega \boldsymbol{\omega}$ and $\dot{\boldsymbol{\theta}}_a^d = -\mathbf{K}_\theta \dot{\boldsymbol{\theta}}$, respectively, where \mathbf{K}_ω and \mathbf{K}_θ are damping gain matrices [44]. Other control designs are also possible, see e.g. [45].

19.3 Modeling and Control of Planetary Robots

Planetary exploration programs have been pursuing extensive scientific missions dedicated to understanding the geological and climatological characteristics of planetary bodies, as well as seeking microorganisms of extraterrestrial life. A robotic probe deployed on a target body plays an important role in achieving scientific missions, in particular, a probe having surface mobility (*rover*) can get close to a specific point of interest and thoroughly enrich the scientific return of the mission.

A fundamental requirement for a rover is the capability of traversing the rough terrain of a planetary body. It also needs to endure a harsh environment: extremely high/low temperatures and/or strong cosmic radiation. A power management scheme for the rover differs from that used for an orbiting (or interplanetary) spacecraft. This is because the power spent by the mobility system significantly varies according to the terrain conditions (sandy, rocky, or sloped terrain) in which the rover travels. The power generated by the solar array panels depends on the solar elevation angle (varied by the local time and latitude of the rover's location) and the orbital longitude of the planetary body.³ The rover should also employ autonomous/semi-autonomous guidance, navigation, and control (GN&C) to travel to a designated location. These technical issues for each subsystem of the planetary rover are summarized in Table 19.1.

From a robotics point of view, this section primarily focuses on the research and development of robotic mobility and GN&C subsystems, and introduces actual applications/implementations of this technology. General descriptions for the other subsystems, including the power, telecommunications, and environmental durability, are presented in other chapters.⁴

The surface mobility system of the rover is indispensable for traversing rough and deformable terrain. Therefore, vehicle/terrain interaction is fundamental mechanics for the following aspects

- *Design*—suspension configuration, vehicle dimensions, and actuator specifications.
- *Mobility evaluation*—slope traversability, obstacle crossing, and power required for the mobility.
- *Navigation and control*—localization, path planning, and traction control.

³ Radioisotope thermoelectric generators (RTG) can solve these limitations for the solar array panels.

⁴ Electrical power is described in Chap. 10, thermal systems in Chap. 13, and telecommunications in Chap. 14.

The surface terrain of the Moon or a planet such as Mars is covered with fine-grained soil (regolith), boulders, rocks, or stones. Because of such challenging terrain, the rover should be aware of mobility hazards such as rolling over a sloped surface, immobilizing wheel slips on loose sand, and colliding with obstacles such as rocks. In particular, the Mars Exploration Rovers (MER), *Spirit* and *Opportunity*, have proven that wheel slip is a critical hindrance to their exploration missions. The issues related to resolving rover mobility requires well-defined mechanics for wheel-terrain interaction and an analytical approach for evaluating rover mobility performance.

The discussion of rover mobility in this section is divided into two issues: the kinematics/dynamics, and the wheel-terrain interaction mechanics. Section 19.3.1 presents the kinematics and dynamics of a planetary rover that can be used for evaluating mobility performance in rough terrain. The wheel-terrain interaction is addressed in Sect. 19.3.2 with a brief review for wheel-terrain interaction research and an introduction to a terramechanics-based analytical model.

The latency in communication owing to the long distance between Earth and a target planet renders the real-time direct teleoperation of a rover infeasible. An operator cannot immediately maneuver the rover when it encounters an obstacle or other contingencies. In addition, the rover cannot obtain prior knowledge of the physical characteristics of an environment. Thus, it needs to consider the environment as it encounters it and make decisions by itself. The GN&C subsystem is designed for these tasks as the autonomous brain of the rover. Section 19.3.3 describes research related to the GN&C, including the sensory system for terrain mapping, localization technique, and path planning.

19.3.1 Kinematics and Dynamics of Mobile Robots

The kinematics and dynamics of a planetary rover are the primary considerations for the mobility analysis of the rover. Whereas there has been work to perfects the kinematics for indoor mobile robots on smooth, flat surface [46–48], the challenge of mobility analysis for a rover is accounting for a rough terrain profile. The motion of the rover becomes relatively complicated because of the dynamic interaction of the wheel on deformable terrain (i.e., wheel slips). The kinematic modeling of a mobile robot on rough terrain has been reported [49–51].

There has also been extensive research regarding the dynamics of planetary rovers: a rover simulator called ROAMS used for the NASA Mars rovers [52], a dynamic

Table 19.1 Technical requirement for rover subsystems

Subsystem	Requirement	Technology
Mobility	Rough terrain traverse (sand, rock, ditch, and crater hill)	Wheels (rigid, inflatable, or flexible), tracked/legged vehicle, suspensions (active/passive)
Power	Power management (generation, distribution, and charging/discharging)	PPT, solar array panel, RTG
Communication	Interplanetary communication TT&C	Antenna/transponder design
GN&C	Terrain mapping, path/motion planning, collision avoidance	Camera, laser range finder, autonomous system
Structures	Launch-proof design, stowed configuration, rover deployment	Launch-lock system, lightweight materials
Environmental durability	Active/passive thermal control, radiation hardening	RHU, heat insulator/dissipation, physical/logical hardening for chip/memory
Mission instruments	Scientific observation, in situ exploration	Robotic manipulation, soil sampling device, subsurface drilling/coring tools, spectroscopic imaging

GN&C guidance, navigation, and control, *PPT* peak power tracking
RTG radioisotope thermoelectric generator, *TT&C* telemetry, tracking, and command
RHU radioisotope heater unit

simulation tool used for ExoMars [53], or a multibody system simulation for a rover on deformable terrain [54, 55].

In this section, the kinematic modeling of an articulated rover on rough terrain is introduced and focused on the inverse kinematics problem and kinematic constraints including wheel/vehicle slips. A dynamic model for the rover is also described.

19.3.1.1 Kinematic Analysis

The kinematics of the rover are basically used for navigation and motion control to achieve appropriate maneuvers on rough terrain. Kinematics also play a significant role in the design perspective: a kinematic model may be used to evaluate joint configuration, link length (between joints), and wheelbase or tread dimensions. In this subsection, an inverse kinematic problem is introduced that can be used to evaluate the kinematic validity and static stability of the rover on rough terrain. Here, a six wheeled rover with a rocker-bogie suspension [56] is assumed for the kinematic analysis. This configuration was used to evaluate the *Sojourner*, MER (*Spirit* and *Opportunity*), and *Curiosity* rovers [57, 58]. In addition, this subsection also addresses a kinematic constraint model for a four-wheeled rover experiencing wheel/vehicle slips. This model can be used for a derivation of the steering maneuver to achieve the desired motion control.

As seen in Fig. 19.11, assuming rover position \mathbf{p}_c and heading Ψ with respect to a terrain given as a height map $z(x, y)$, the kinematic loop closure equations can be written as follows [59]

$$\begin{aligned}
 z_{rr} &= z_{lr} + l_1 \cos \Theta (\sin \theta_{1r} - \sin \theta_{1l}) + w \sin \Theta \\
 z_{rr} &= z_{lm} + \cos \Theta (l_1 \sin \theta_{1r} - l_2 \sin \theta_{1l} - l_3 \sin \theta_{2l}) + w \sin \Theta \\
 z_{rr} &= z_{lf} + \cos \Theta (l_1 \sin \theta_{1r} - l_2 \sin \theta_{1l} - l_4 \sin \theta_{2l}) + w \sin \Theta \\
 z_{rr} &= z_{rm} + \cos \Theta (l_1 \sin \theta_{1r} - l_2 \sin \theta_{1l} - l_3 \sin \theta_{2r}) \\
 z_{rr} &= z_{rf} + \cos \Theta (l_1 \sin \theta_{1r} - l_2 \sin \theta_{1l} - l_4 \sin \theta_{2r})
 \end{aligned} \tag{19.37}$$

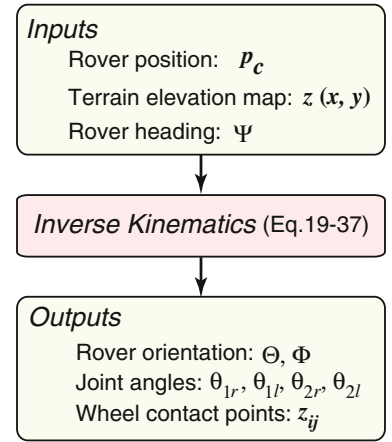
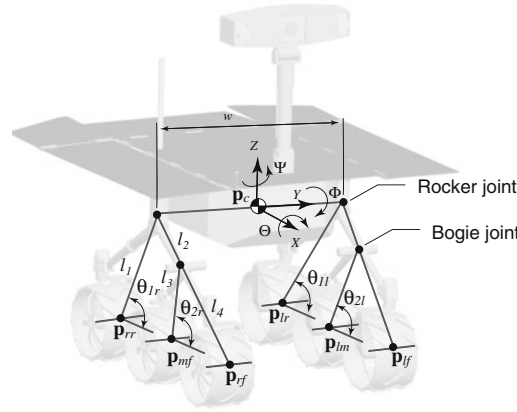
where z_{ij} ($i = \{r, l\}$, $j = \{r, m, f\}$) refers to the z component of \mathbf{p}_{ij} , with index i referring to the right and left side, and index j referring to the rear, middle, and front wheels.

Inputs for this equation are a terrain elevation map, the position \mathbf{p}_c of the rover center, and the rover heading. These inputs mitigate the number of unknown parameters that can be determined by solving the equation. The solution for the inverse kinematic problem with multiple contact points on the terrain is subject to the simultaneous cross-solution of multiple nonlinear equations. Newton's method can be applied to solve such equations.

The kinematics of the rover is also used for motion control of the rover, such as the steering maneuvers needed to follow a specified traveling path. As mentioned in Sect. 19.3.1, wheel/vehicle slips are a critical issue for the rover; therefore, the kinematic model for motion control should include such effects. The rest of this subsection describes the kinematic model with wheel/vehicle slips [60].

A 2D kinematic model of a four-wheeled vehicle, which includes the slip angle of the vehicle β_0 and lateral wheel slippage β_i is shown in Fig. 19.12. In this model, each wheel has a steering angle δ_i , where the subscript i denotes the wheel ID ($i = 1, \dots, 4$, in this case). The position and orientation of the centroid of the vehicle defined as (x_0, y_0, θ_0) ,

Fig. 19.11 Kinematic description of six-wheeled rover with a rocker-bogie suspension



and (x_i, y_i) give the position of each wheel. The dimension of the rover is defined by $l_f, l_r, d_R,$ and d_L . For this model, the following assumptions are considered: (1) the distance between wheels is constant, (2) the steering axle of each wheel is perpendicular to the terrain surface, and (3) the vehicle does not consist of any flexible parts.

The non-holonomic constraints with the lateral slips of the wheel and vehicle are defined by the following equations

$$\begin{aligned} \dot{x}_0 \sin \phi_0 - \dot{y}_0 \cos \phi_0 &= 0 \\ \dot{x}_i \sin \phi_i - \dot{y}_i \cos \phi_i &= 0 \end{aligned} \quad (19.38)$$

where $\phi_0 = \theta_0 + \beta_0$, and $\phi_i = \theta_0 + \delta_i + \beta_i$. The geometric constraints between the centroid of the vehicle and each wheel are written as

$$\left. \begin{aligned} x_1 &= x_0 + l_f \cos \theta_0 - d_L \sin \theta_0 \\ x_2 &= x_0 - l_r \cos \theta_0 - d_L \sin \theta_0 \\ x_3 &= x_0 - l_r \cos \theta_0 + d_R \sin \theta_0 \\ x_4 &= x_0 + l_f \cos \theta_0 + d_R \sin \theta_0 \end{aligned} \right\} \rightarrow x_i = x_0 + X_i \quad (19.39)$$

$$\left. \begin{aligned} y_1 &= y_0 + l_f \sin \theta_0 + d_L \cos \theta_0 \\ y_2 &= y_0 - l_r \sin \theta_0 + d_L \cos \theta_0 \\ y_3 &= y_0 - l_r \sin \theta_0 - d_R \cos \theta_0 \\ y_4 &= y_0 + l_f \sin \theta_0 - d_R \cos \theta_0 \end{aligned} \right\} \rightarrow y_i = y_0 + Y_i. \quad (19.40)$$

Given the desired heading angle $\theta_0 = \theta_d$ and desired linear velocity v_d , the desired steering maneuver (i.e. steering angle δ_i) is elaborated as follows: first, transform Eq. 19.38

$$\delta_{di} = \tan^{-1}(\dot{y}_i/\dot{x}_i) - \theta_d - \beta_i \quad (19.41)$$

and then, substitute Eqs. 19.39 and 19.40 into Eq. 19.41. The desired steering angle is determined as follows

$$\delta_{di} = \tan^{-1} \left(\frac{v_d \sin \theta_d - \dot{Y}_i(\dot{\theta}_d)}{v_d \cos \theta_d - \dot{X}_i(\dot{\theta}_d)} \right) - \theta_d - \beta_i. \quad (19.42)$$

The desired velocity v_d and heading angle are derived based on a path following control strategy such as the pure-

pursuit algorithm [61], or path following control with slip compensation [60, 62].

19.3.1.2 Dynamic Analysis

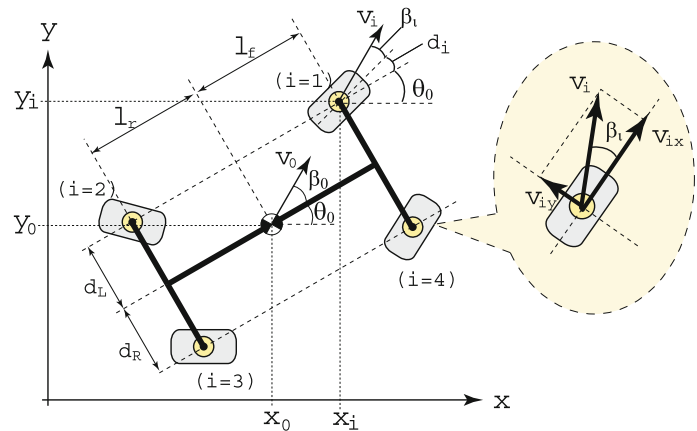
The motion profile of the entire rover can be numerically evaluated by using a dynamic model. Despite the slow traveling velocity of a rover,⁵ the motion often behaves dynamically because of rough terrain such as bumpy, sloped, or rocky surfaces. A schematic illustration of the dynamic model of a six-wheeled rover having a rocker-bogie suspension is shown in Fig. 19.13. The dynamics of the rover are modeled as an articulated multibody system as follows [63]

$$\mathbf{H} \begin{bmatrix} \dot{\mathbf{v}}_b \\ \dot{\mathbf{q}} \end{bmatrix} + \mathbf{C} + \mathbf{G} = \begin{bmatrix} \mathcal{F}_l \\ \tau \end{bmatrix} + \mathbf{J}^T \mathcal{F}_e \quad (19.43)$$

where \mathbf{H} represents the inertia matrix of each body, \mathbf{C} is the velocity depending term, \mathbf{G} is the gravity term, \mathbf{v}_l are the translational and angular velocities of the vehicle, \mathbf{q} is the angle of each joint (such as wheel rotation and steering angle), \mathcal{F}_l are the forces and moments at the centroid of the vehicle body, τ are the torques acting at each joint (driving/steering torques), \mathbf{J} is the Jacobian matrix, and \mathcal{F}_e consists of the external forces and moments acting at the centroid of each wheel, namely $f_{ij}(i = \{r, l\}, j = \{r, m, f\})$. The external (contact) forces and torques on each wheel can be calculated based on a wheel-terrain contact model, as described in the next section. The dynamics of a rover for given traveling and steering conditions are numerically obtained by successively solving Eq. 19.43.

⁵ The average velocity of an MER was about 0.01 m/s. The Mars Science Laboratory (MSL) *Curiosity* was designed to travel up to approximately 200 m per day [64].

Fig. 19.12 Kinematic model of four-wheeled rover with wheel/vehicle slips



19.3.2 Wheel-Terrain Interaction Mechanics

The study of the mechanical properties of the terrain and the terrain's response to an off-road vehicle has been included in the field of terramechanics,⁶ in which an analysis of the interaction between wheel/track and soil has been of primary focus.

In classical terramechanics, Bekker, an originator of terramechanics, derived a well-known pressure-sinkage equation and also formulated the shear stress as a function of soil deformation (displacement) [65, 66]. His work greatly contributed to the design and development of the Lunar Roving Vehicle used on the Apollo 15–17 missions to the Moon. Wong developed a comprehensive procedure for predicting the performance of both driven and towed wheels [67–69]. The procedure calculates wheel mechanics by applying the stress distribution model beneath the wheel.

Terramechanics can be divided into three methods [70, 71]: (1) an analytical method, (2) an empirical method, and (3) a numerical method.

The analytical method considers a physical model for vehicle-terrain interactions based on a theoretical analysis with experimental results for model validation. The empirical method uses a practical measurement of soil strength with a specialized apparatus, such as a cone index (CI) [67], which is often used for an in situ prediction of vehicle traversability. The numerical method includes the finite-element method and discrete-element method that simulate soil deformation and vehicle-terrain interaction behavior with computer technology [72–74].

The wheel-terrain model can be used for the design of rover mobility systems: the terramechanics model can be used as a feasible wheel/track design because it is able to maximize the traction performance for off-the-load locomotion under specific constraints [75, 76]. Additionally, the

mobility performance of the rover (i.e., its traversability on sloped or deformable terrain) will be numerically/experimentally analyzed based on the wheel model [77, 78]. This mobility prediction and evaluation technique would be also valuable for the mobility system design [79] in addition to an actual rover operation to determine rover maneuvering. Some recent works have reported dynamic simulation tools combined with the terramechanics wheel model (e.g., NASA Mars rovers [52, 80] and ExoMars [53, 55]).

This section focuses on the analytical method and introduces a typical interaction model of a rigid wheel on deformable terrain.

19.3.2.1 Terramechanics-Based Wheel-Terrain Model: Analytical Method

In the analytical method, the basic principle of a wheel traction model considers the stress distribution at the wheel-terrain contact point, which usually depends on wheel slips. An integral of the stress around the contact point derives wheel traction forces, such as drawbar pull, side force, and resistance torque.

A contact model for a rigid wheel on deformable terrain is schematically shown in Fig. 19.14. A classical terramechanics model defines the wheel-terrain contact forces, including the drawbar pull F_x , vertical force F_z , and resistance torque T , as the following equations [68]

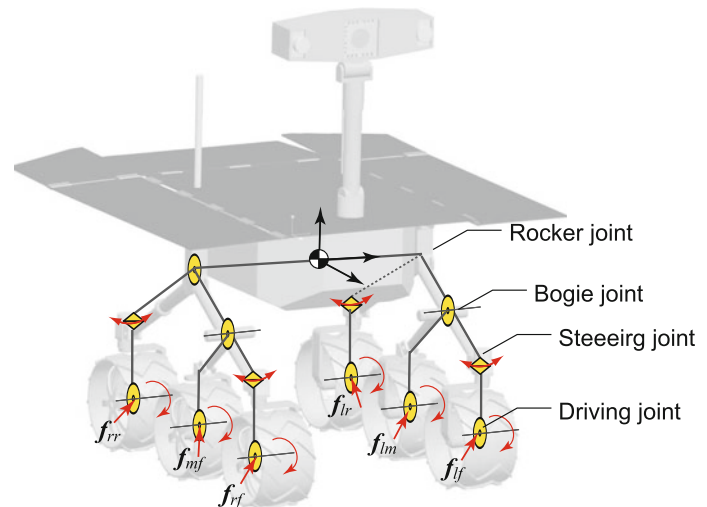
$$F_x = rb \int_{\theta_r}^{\theta_f} \{\tau_x(\theta) \cos \theta - \sigma(\theta) \sin \theta\} d\theta \quad (19.44)$$

$$F_z = rb \int_{\theta_r}^{\theta_f} \{\tau_x(\theta) \sin \theta + \sigma(\theta) \cos \theta\} d\theta \quad (19.45)$$

$$T_x = r^2 b \int_{\theta_r}^{\theta_f} \tau_x(\theta) d\theta \quad (19.46)$$

⁶ The term terramechanics is coined from 'terrain' and 'mechanics'. Soil mechanics is the study of the interaction of structures in various soils.

Fig. 19.13 Rover dynamics model



where b represents the wheel width, $\sigma(\theta)$ is the normal stress beneath the wheel, and $\tau_x(\theta)$ are the shear stresses in the longitudinal direction of the wheel. The contact point of the wheel is determined by the entry angle θ_f and the exit angle θ_r .

The side force (i.e., the force in the lateral direction) of the wheel appears when the wheel steers or traverses sloped terrain. The side force F_y can be modeled as the summation of two forces generated at the wheel: the force F_u attributable to the shearing motion beneath the wheel and the force F_s generated by the bulldozing motion on the side face of the wheel [63]

$$F_y = F_u + F_s = \int_{\theta_r}^{\theta_f} rb\tau_y(\theta) + \int_{\theta_r}^{\theta_f} R_b\{r - z(\theta) \cos \theta\}d\theta \quad (19.47)$$

where $\tau_y(\theta)$ are the shear stresses in the lateral direction of the wheel and R_b is modeled as a reaction resistance generated by the bulldozing phenomenon on a side wall of the wheel. R_b is a function of the wheel sinkage z .

In these equations, the normal stress $\sigma(\theta)$ and shear stresses $\tau_x(\theta)$ and $\tau_y(\theta)$ are defined by the function of soil parameters, wheel contact angle, and wheel dimensions. Details about the stress model can be found in other research [63, 68, 69, 81, 82].

19.3.2.2 Experimental Validation

The previous wheel traction model needs to be validated through multiple experimental tests with varied state parameters such as soil or wheel traveling profiles. A single-wheel test bed (Fig. 19.15) is commonly used for model validation. The test bed primarily consists of a carriage section and wheel section. The carriage velocity is

controlled relative to wheel velocity, which realizes wheel slip (or traction load), while measuring wheel traction forces, wheel sinkage, and other parameters. Experimental data are then compared with the values obtained from the numerical simulation of the wheel traction model.

The primary focus of the classical terramechanics model has been devoted to the application of large, heavy vehicles (i.e., vehicles weighing hundreds/thousands kilograms). Therefore, when exploiting the classical model for analyzing lunar/planetary rover test beds (usually small, lightweight), several assumptions for the classical model would be omitted⁷ that may cause an inaccurate calculation of wheel traction performance.^{8,9} Some researchers have assumed the errors attributable to the omitted assumptions as modeling errors or the uncertainty of parameters used for the calculation. Recently, several approaches to update/improve the classical terramechanics model were successfully applied to relatively lightweight vehicles. For example, a direct measurement device for the normal stress distribution has been reported [84]. A wheel-diameter dependent pressure-sinkage model has been proposed [85]. An improved approach for the calculation of shear deformation modulus has also been studied [86].

⁷ One assumption in Bekker's pressure-sinkage model is that the contact point of the wheel on deformable soil (circumferential section) is a series of consecutive flat plates.

⁸ Bekker noted this issue: "Predictions for wheels smaller than 20 inches in diameter become less accurate as wheel diameter decreases, because the sharp curvature of the loading area was neither considered in its entirety nor is it reflected in bevameter tests" [66].

⁹ These assumptions provide an inaccurate prediction for vehicles with wheel diameters less than approximately 50 cm and a normal loading of less than approximately 45 N [85].

Fig. 19.14 Wheel-terrain contact model

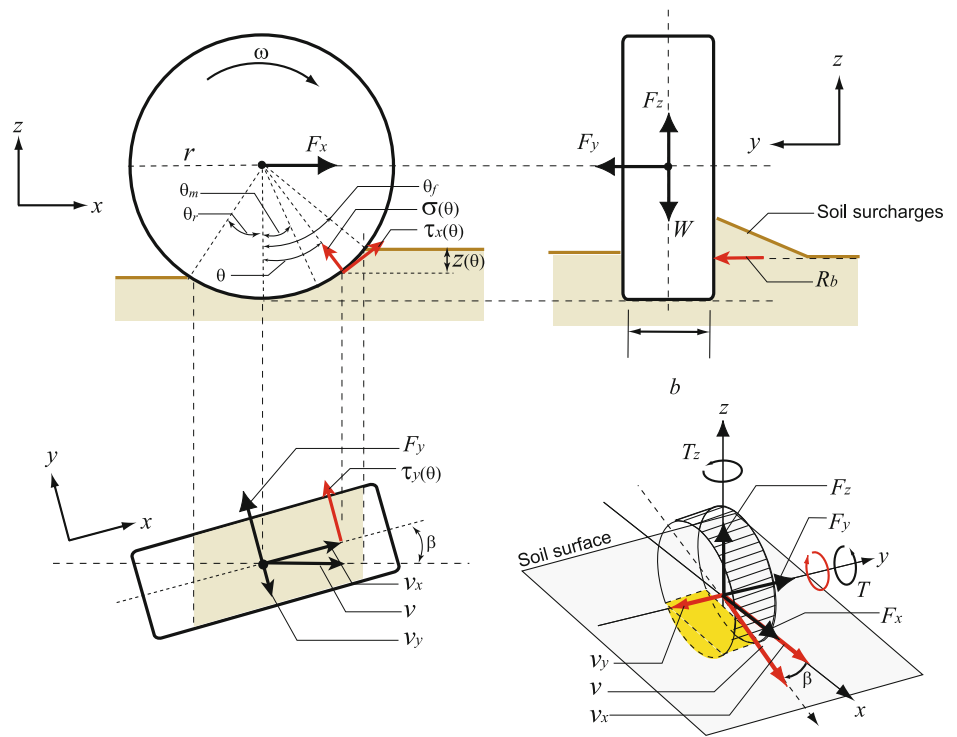
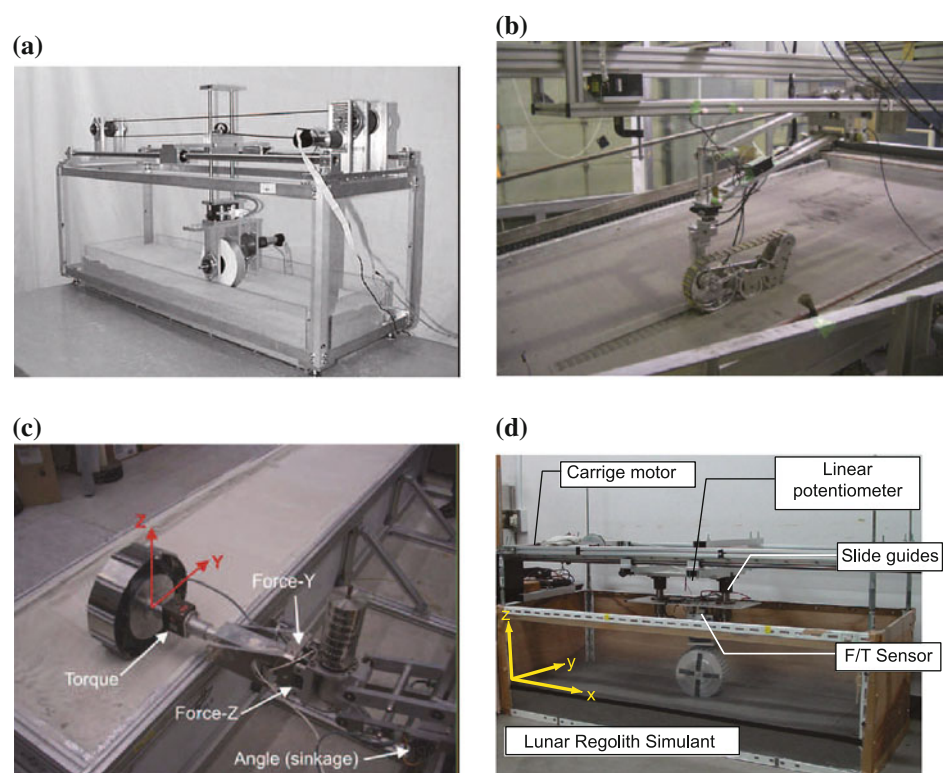


Fig. 19.15 Single-wheel test beds for experimental validation of terramechanics models. **a** Single-wheel test bed at MIT [83]. **b** Single-track test bed at JAXA [75]. **c** Single-wheel test bed at DLR [76]. **d** Single-wheel test bed at Tohoku University [63]



19.3.2.3 Soil Parameter Identification and its Uncertainty Analysis

The wheel-terrain interaction model described in the previous section assumes that the physical properties of the soil

are known. These properties must be measured in situ by on-board robotic sensor systems [87], but their values would stochastically vary with location, resulting in tremendous uncertainties.

Several researchers have addressed soil parameter identification, for example an online terrain parameter estimator that uses a linear least-squares method to compute the values of cohesion and internal friction angle with simplified classical terramechanics equations [88], and applying the Newton–Raphson method to a modified nonlinear wheel-terrain interaction model that can identify unknown parameters such as the pressure-sinkage coefficient, internal friction angle, and shear deformation modulus [89].

The parameters identified by these approaches remain subject to uncertainty. Some recent works have attempted to predict rover mobility even under uncertain conditions, for example a learning-based approach for slip prediction that is used for a traversability analysis of a rover [90], and an applied a statistical method for mobility prediction that explicitly considers terrain uncertainty and achieves a computationally-efficient prediction of rover dynamics [91].

19.3.3 Guidance, Navigation, and Control

Planetary rovers need to traverse the surface of a target body with little knowledge of the terrain, such as the physical properties of the soil or the geometrical features of the terrain. Space probes and orbiters around the target body may be able to provide a global terrain map with relatively good accuracy.¹⁰ The terrain map available from the orbiter is often useful for determining a ‘global’ destination; however, it is not feasible to refer to the map in real-time while the rover travels through intermediate waypoints toward the global destination. Therefore, the rover is required to perceive the local terrain environment and to plan a feasible path to traverse rough terrain. This section introduces the research and development dedicated to terrain mapping, rover localization, and path planning; these are key techniques for the GN&C systems of the rover.

19.3.3.1 Terrain Mapping

Once a rover is deployed on a planetary body, it must first measure terrain features (terrain mapping). 3D information from the terrain map can be exploited to assess obstacle size, slope angle, or terrain roughness so that the rover can plan the path to travel on the map. In addition, an augmented map of the terrain environment can be generated from consecutive maps.

Stereo vision (i.e., visual information taken by a stereoscopic camera mounted on the rover) is a particular technique by which to obtain 3D terrain mapping [92, 96]. An example of the stereo vision results from a MER (*Spirit*)

is shown in Fig. 19.16 [94]. The bottom image of the figure shows an elevation plot of the scene taken from stereo cameras.

Sufficient progress in terms of radiation-hardened flight CPUs for space probes in the last few decades accelerated the on-board stereo vision process, but stereo camera-based terrain mapping is still a time-consuming task for the low-power CPU on the rover because stereo images should be correlated to one another by stereo matching, thus requiring a relatively long computational time [94]. Also, the visual information provided by the camera may vary with the intensity of sunlight.

Another technique for terrain mapping is the use of a laser range-finder (LRF) or laser imaging detection and ranging (LIDAR) that can determine the distance from a laser emitter to an object based on the time-of-flight principle. There has been extensive research and development in which the LIDAR technique was used in robotics for sensing the environment and for classifying the terrain [97, 98]. In particular, the Defense Advanced Research Projects Agency (DARPA) Ground Challenge and Urban Challenge programs have accelerated the development of LIDAR and its implementation for robotic mobile vehicles [99, 100]. Figure 19.17 represents an example of LIDAR-based terrain mapping.

Although a space-hardened LIDAR was used for the rendezvous and docking of the Space Shuttle to the International Space Station [101, 102], as of 2012, no actual rover has been equipped with LIDAR. Several research and development efforts have been reported that introduce LIDAR techniques and applications for a rover [103–105].

The LRF can measure 3D distances from the sensor to objects, providing a ‘point cloud’ of data of the scene without additional processes (c.f., camera-based mapping needs stereo matching for the 3D mapping). A drawback of the LIDAR sensor is that the scanning mechanism including the actuators and their movable parts may be less durable during launch vibrations and/or landing shocks. Alternatively, as a solid-state LIDAR sensor, a 3D flash LIDAR imaging system, is being developed that can capture the real-time 3D depth and intensity of a scene. The flash LIDAR consists of CMOS-based avalanche photodiode detectors, each pixel of which enables the measurement of the range and intensity of the light illuminated by the laser. Therefore, the flash LIDAR acts like a 2D image-plus-depth camera that achieves the relatively fast capturing of the terrain without any movable parts and actuators.

19.3.3.2 Localization

A rover needs to measure and update its position and orientation during its travel on the map obtained. An accurate measurement of position and orientation is challenging because the globally aided navigation schemes, such as the

¹⁰ The Mars Reconnaissance Orbiter launched by NASA achieved 0.3 m resolution with a high-resolution imaging science experiment (HiRISE) camera.

Fig. 19.16 MER (*Spirit*)
hazcams stereo imagery results
(from [94])

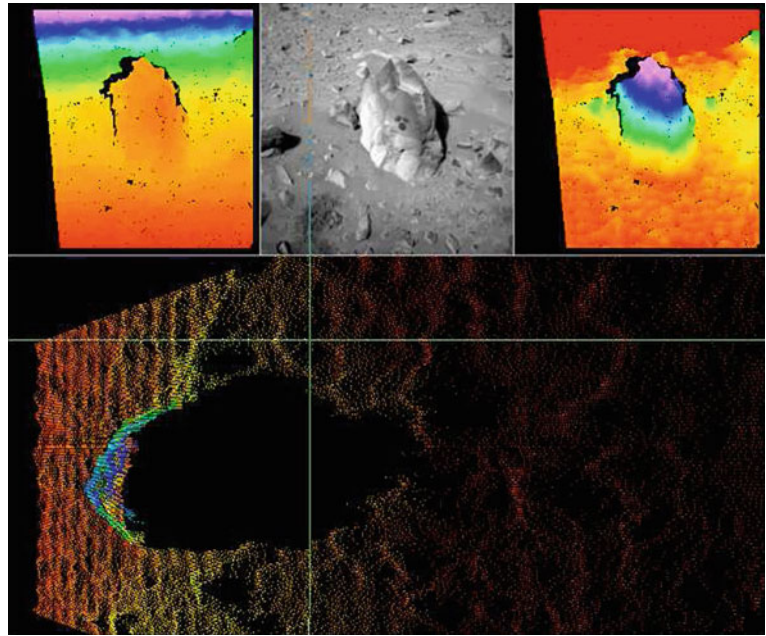
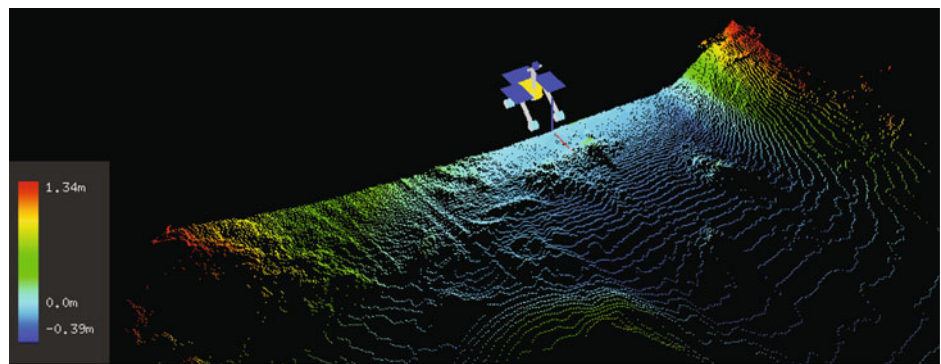


Fig. 19.17 LIDAR-based
terrain mapping result (from
[126])



global positioning system (GPS), or heading reference relative to a global magnetic field is not available on planetary bodies.

The internal state sensors such as an inertial measurement unit (IMU) and wheel encoders are often used to achieve a position/pose estimation of the rover by dead reckoning. A sophisticated estimate method with a Kalman filter may be applied to reduce measurement noise. A pose estimation method using stereo imagery with learning from previous examples of traversing similar terrain was proposed by [106, 107]. The MERs have exploited Sun sensing with their cameras for occasional heading updates [108].

Odometry using wheel encoders is a traditional approach to measuring distance traveled; however, it may not be reliable if the rover travels on sandy loose terrain in which the wheels slip, resulting in incorrect calculations of distance traveled with respect to wheel rotations. The errors accumulate over time and will degrade the accuracy of the position estimation. To resolve this drawback, image-based

odometry, termed visual odometry, has been widely applied to planetary rovers [108–110]. Visual odometry estimates the traveling velocity of the vehicle using the optical flow vectors between the time-consecutive images taken by an on-board camera(s). Integrating the velocity estimates with IMU readouts or stereo images for pose estimation provides an accurate estimation of the six degrees of freedom of the rover's motion. The visual odometry system of the MER was used for more than 14 %¹¹ of its first 10.7 km of travel [110].

For the MERs, a bundle adjustment technique was implemented to update and correct rover localization. The technique uses a stereo pair image and manually selected tie-points on the images to create a geometric configuration of the image. The accumulated images taken day by day

¹¹ A high computational burden is the reason for such a short usage of the visual odometry.

propagate the entire image network and determine the global position of the rover on the map [108].

19.3.3.3 Path Planning

The latency in communicating between Earth and the rover on a planetary body often impedes direct teleoperation; therefore, the rover must possess a high degree of autonomous mobility for traversing unknown rough terrain. One primary task for such autonomy is to find a feasible path on the map generated by the on-board sensors and to avoid mobility hazards.

Substantial works dedicated to the path/motion planning of mobile robots have been performed, such as the A* and D* methods [111], the potential field approach [112], the probabilistic roadmap technique [113], and the rapidly exploring random tree (RRT) algorithm [114]. Randomized approaches to kinodynamic motion planning [115] have been reported to be an efficient tool for the purpose of path generation, with RRTs proving to be a highly effective framework. Also, a heuristically biased expansion for generating efficient paths that satisfy dynamic constraints has been developed by [116]. Explicit modeling of a robot's closed-loop controller in the planning method, which results in trackable paths, has also been studied [117].

Robotic mobility in path planning is important for field conditions in which terrain inclination, roughness, and mechanical properties can significantly degrade a rover's mobility. Path generation techniques that consider robotic mobility have also been investigated. For example, a trajectory generation method on rough terrain, accounting for predictable vehicle dynamics, has been proposed [118]. A planning algorithm with model-based evaluations, which include the uncertainties of terrain measurement and rover localization, has been developed [59]. In addition, a terrain traversability index with fuzzy logic for mobile robot navigation has been introduced [119], and its terrain traversability map has been used for the path planning of planetary rovers [120]. An explicit consideration of the dynamic mobility of a rover in path planning and an energy-based evaluation of candidate paths has been proposed [121], see Fig. 19.18.

The MERs have autonomous navigation with hazard avoidance technology based on a local path planner called GESTALT (i.e., grid-based estimation of surface traversability applied to local terrain, see Fig. 19.19) [93, 122]. The local terrain map created by the on-board stereo camera pair is a grid-based map, with each grid containing a goodness value indicating the terrain traversability. Then, several candidate trajectories, including forward and backward arcs, and two-point turns are evaluated. The trajectory that has the best goodness value is chosen and then the rover executes the predetermined distance and trajectory. The flight software of the MER has been upgraded to manage

conflict voting between hazard avoidance and waypoint selection, achieving simultaneous local and global path planning with the Field D* algorithm [123–125].

19.4 Telerobotics

Telerobotics is a technology developed for the remote control of space robots. The primary purpose is the handling of the communication time delays that occur during teleoperation from the ground to a robot in orbit or on the Moon. A communication time delay of 4–7 s usually occurs in such teleoperation, which is the inherent time lag that affects most communications equipment used for transmitting telemetry data. In teleoperation between the Earth and Mars, for example, the time delay is as much as several minutes and is largely dependent on the distance.¹² This forces the operator to adopt a move-and-wait strategy in executing remote tasks. The operator has to await the response and check it with each command sent. Accordingly, the extended communication time delay reduces efficiency and increases waiting time [127].

Meanwhile, a hierarchical structure can be seen in the task shown in Fig. 19.20. A higher-level (complex) task comprises multiple lower-level (simpler) tasks and this pyramid structure relates to the level of autonomy. Upper-level tasks require a higher level of autonomy. From the perspective of the operator-robot relationship, higher-level commands can reduce command frequency, and consequently the checking frequency and the overall waiting time. Accordingly, higher-level autonomy ease the adverse effects of the communication time delay. This is a basic concept of telerobotics and a standard framework for space teleoperation.

Conversely, direct teleoperation by means of a joystick is a typical example of the use of lower-level commands. Generally, such systems are significantly affected by communication time delay. However, joystick systems are one of the key framework elements of teleoperation, including space teleoperation, since short-distance teleoperation from

¹² The latency is a summation of the propagation time of the radio wave and the delays of signal processing in the computers and communications nodes. For example, in case of the ISS at 400 km altitude, the direct round-trip radio-propagation delay is just 0.003 s. But if the communication is linked via a geostationary satellite at 36,000 km altitude, the round trip delay increases to 0.5 s. The ETS-VII, which was a low-Earth orbit satellite at about 550 km altitude, utilized the round-trip of two different geostationary satellites and, with cumulative delays in the transmission nodes, the total latency was 5–6 s in practice. Between the Earth and the Moon, the round-trip delay due to just the distance is 2.5 s. For the Mars, it varies from 6.2 to 45 min depending on the relative positions of Earth and Mars in their orbits.

Fig. 19.18 Path planning and evaluation simulation (from [121])

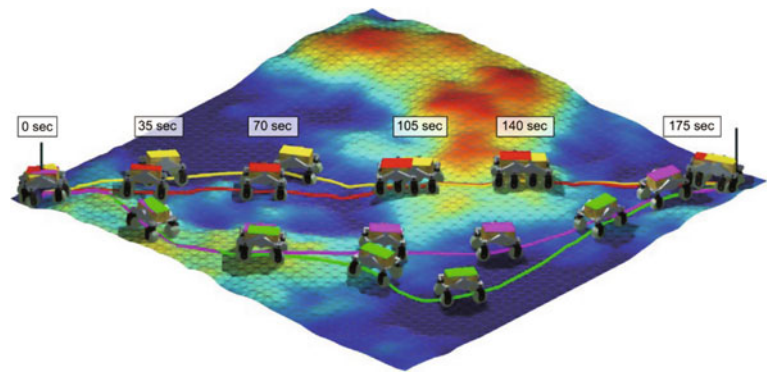
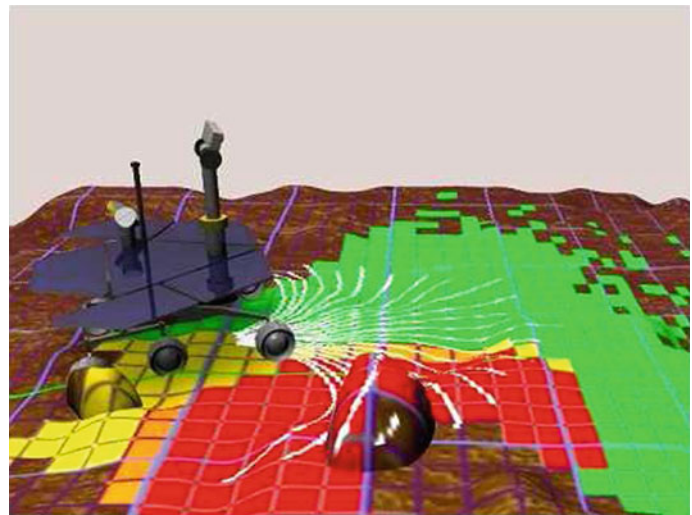


Fig. 19.19 Illustration of terrain assessment and path selection. *Red* cells indicate unsafe areas around the large rock, *yellow* cells indicate traversable but rougher areas around the smaller rock, and *green* cells indicate safe and flat areas (from [122])



cabins are not subject to serious communication time delays. Direct teleoperation is part of telerobotics.

19.4.1 Direct Teleoperation

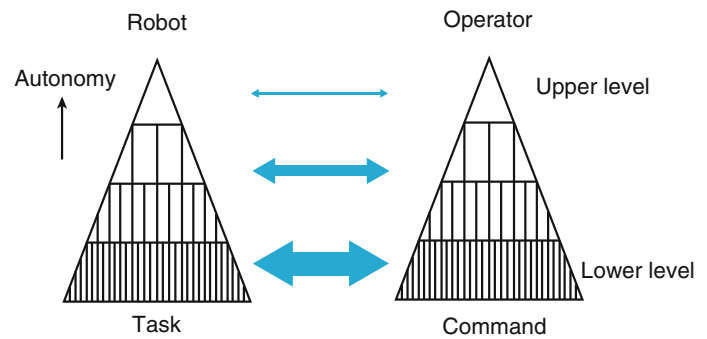
Direct teleoperation utilizes continuous low-level commands, e.g. the position or velocity of the end-effector, and includes control approaches. In unilateral control, the operator commands the position or velocity of the end-effector, but the motions of the remote robot are not signaled to the operator except for visual information. The joystick is the most popular input device for unilateral control. Meanwhile, a master–slave manipulator system is utilized for bilateral control. The master arm is an input device, while the slave arm is a remote manipulator. The master arm can display both the motion and force of the slave arm. The force information is very useful in undertaking skillful tasks. However, bilateral control is not part of mainstream space teleoperation, because it is significantly affected by communication time delay. To date, only a few advanced experiments of bilateral control between the ground and robots in orbit have been performed.

19.4.2 Unilateral Control

Rate control is the most popular approach to unilateral control when teleoperating a robot with a joystick. The SRMS (Shuttle Remote Manipulator System) also employs rate control with joysticks in the cabin. In the SRMS, two joysticks named the Translational Hand Controller (THC) and Rotational Hand Controller (RHC) are used for translational and rotational motions, respectively, as shown in Fig. 19.21. A 6-axis joystick, e.g. SpaceMouse by 3D connection, is also available on the ground, but the combination of THC and RHC has become the standard space application input device due to its long history. Astronauts in particular prefer this combination, because they have extensively trained with the devices for extended periods. Both the JEMRMS (Japan Experiment Module Remote Manipulator System) and Canadarm2 also employed these two joysticks.

As noted earlier, communication time delay is a critical issue in space teleoperation from the ground. Predictive display was introduced in [128] to address this. The predictive display function indicates the future position of the manipulator by computer graphics, whereupon the operator can

Fig. 19.20 Task level and command level



teleoperate the remote manipulator as if there were no time delay. Accordingly, the predictive display improves operational efficiency, even for low-level commands, as the operator can continuously send commands that resemble higher-level commands but include a range of lower levels ones. This reduces the checking frequency required and mitigates the adverse effects of the communication time delay.

There have been very few attempts at direct teleoperation from the ground involving real space robots in orbit. ROTEX (Robot Technology Experiment), developed by DLR, achieved the first direct teleoperation from the ground [129]. This involved a 6-axis Space Ball employed as the input device, whereby a precise simulator in a ground-based workstation that predicted the robot motion and the environment in which to compensate for the communication time delay. The simulator included both geometrical and dynamic models. It predicted the motions of a floating object. The ETS-VII (Engineering Test Satellite No. 7) developed by NASDA (currently JAXA) also achieved direct teleoperation from the ground by joysticks and rate control [130].

In practice, rate commands are integrated on the ground, and the results are sent in the form of positional information to the remote robot in orbit, which comprehensively protects its motion when the communications link is broken.

19.4.3 Bilateral Control

Bilateral control is achieved by a master–slave manipulator system. Initially, a master–slave manipulator with the same structure and DOF was employed. Currently however, a different structural master arm is often used, because the motion of the end-effector is a critical issue. It should be noted that if the slave arm has redundant motion, an additional approach is required to operate the redundant joint with a different structural master arm. Through the master–slave manipulator, the operator can sense both the motion and force at the remote site. Although, the slave arm executes the force of the operator, the communication time delay makes some bilateral controls impossible. In response, [131] introduced a scattering transformation approach that ensures system stability. However, the master arm must have

a heavier operational feeling to ensure stability, given the extended communication time delay. In practice, the acceptable limit for communication time delays is less than one second, which means that bilateral control cannot be used for Earth-based teleoperation of robots in orbit which entails a communication time delay of several seconds.

A few attempts at master–slave control of a real orbital robot have been made. ETS-VII carried out experiments with a master arm [132], in which bilateral control was locally achieved by means of a virtual model on the ground. The reference position, based on the reference force exerted by the operator, was sent to the slave arm, which executed the reference force by compliance control. The remote environment should be known in such a process. Furthermore, real bilateral control in a large loop, that includes the ground and the orbit was also executed on ETS-VII [133]. The operator could feel the remote force with a communication time delay of almost 7 s, but it was difficult to apply the approach to practical tasks as mentioned above. Meanwhile, the ROKVISS (Robotics Component Verification on ISS) developed by DLR also achieved bilateral control [134]. In this project, a round trip delay of 10–20 ms was achieved, because the operator site on the ground was directly connected to the ISS, making reasonable bilateral control possible.

19.4.4 Supervisory Control

Supervisory Control is a concept proposed by Sheridan which includes not only telerobotics but also various semi-autonomous systems [135, 136]. The term ‘supervisory control’ has a longer history than that of telerobotics and establishes a framework for the relationship between humans and semi-autonomous systems. Basically, humans issue higher-level commands and monitor the results as supervisors, while semi-autonomous systems execute the commands as subordinates. Similar relationships can be found, not only in space robots but also various other systems. Fig. 19.22 shows a typical example of supervisory control in a space robot system. The robot achieves semi-autonomous functions with local loops based on various

Fig. 19.21 Astronaut Leroy Chiao, expedition 10 commander and NASA ISS science officer, works with the controls of the Canadarm2, or space station remote manipulator system (SSRMS) in the Destiny laboratory of the International Space Station (18 October 2004).
Image NASA



sensors. On the control site, the operator sends commands via a computer-assisted Human Interface (HI). The Human Interactive Computer (HIC) includes a model of the remote environment and an expert advisory system, based on prior information. The HIC also interacts with the operator through sensors and actuators. Autonomy on the HI side is therefore also important. In ROTEX, a multisensory gripper that included various sensors was a key technology for achieving good performance. Intelligent sensory feedback capabilities compensate for errors that the predictive graphic simulator cannot handle.

19.4.5 Relationship Between Humans and Systems

Ensuring a reasonable relationship between humans and systems depends on both applications and the current level of technology. The first question that must be asked is whether humans always maintain superior positions to systems. Supervisory control clearly depicts humans acting as supervisors and making the final decisions. Shared control and traded control however, show different frameworks afford flat relationships. Humans perform the tasks to which they are best suited, and robots also do likewise in accomplishing difficult tasks that cannot be achieved without assistance. In shared control, a task is simultaneously shared between a human and a robot. For example, the human controls the trajectory of the end-effector in grasping a glass full of water, while the robot keeps the water from spilling. Task sharing is a key feature of shared control. Conversely, in traded control, humans and robots work in turn, which means the tasks are divided by time. For example, a human firstly decides on the path plan, whereupon the robot checks for possible collisions with obstacles. Alternation timing is a major aspect of in traded control.

The relationship between humans and robots is a subject of debate, not only in space robotics, but also in the human factors in the U.S. and ergonomics in Europe. Human factors research started by analyzing airplane accidents that occurred during World War II. Currently, both words are used for the same meaning. These fields show the value of enhancing safety.

In Germany, the 30 min rule is well known for nuclear power plants. In emergencies, the system should handle all trouble during the first 30 min. In other words, the human operator should not intervene in the operation during this period, but instead gather information and prepare the best solution. This protects against human errors caused by panic and is made possible by the slow process of nuclear power plants. It is noteworthy that during the first 30 min the system adopts a superordinate stance compared to that of humans.

Conversely, in shared control, there is the potential for the actions of humans to conflict with those of robots. The operator should recognize what is happening in the system, otherwise a serious accident may occur. Regardless of circumstances, the relationship between humans and systems should be designed to avoid human errors. In space robots, serious failure is unacceptable due to the cost involved, while safety for astronauts is paramount. The scope of activities in space is expanding to include work in orbit, on the Moon, on Mars and beyond. More critical work would be necessary, which would require the establishment of a proper relationship between humans and systems.

19.4.6 Human Interface

The operator teleoperates a remote robot via a human interface. An intuitive and easily understandable human interface should be provided. A wire-frame graphic model may be superimposed on a real video image to show a predictive display, as in [137]. Conversely, the real video

Fig. 19.22 An example of supervisory control

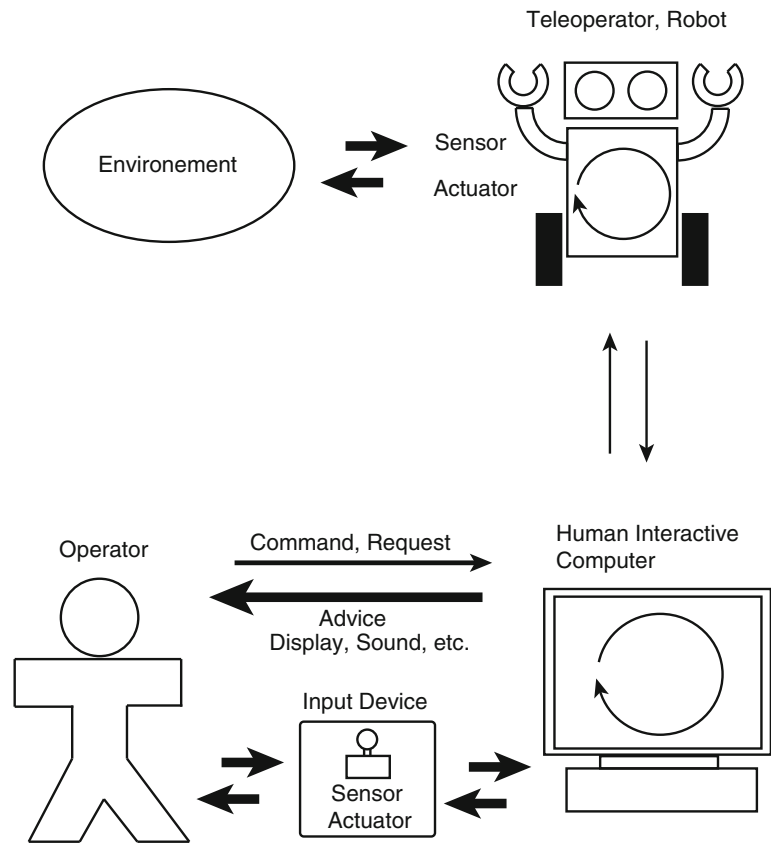


Fig. 19.23 Human interface for Robonaut 2



image is installed into the 3D graphic model as texture to understand the camera posture in [138]. It is therefore important to display incomprehensible invisible information, for which a multi-modal interface, including voice, is a key technology. For Robonaut-2 a novel interface was developed where the motions of the operator are captured by a motion tracking system and a head-mounted display is employed to enhance presence; see Fig. 19.23. Robonaut-2 directly follows human motions but includes an indexing function because of the difference in size. This indexing allows each motion to be connected and disconnected with

an offset, which means the operator can intuitively teleoperate Robonaut-2. The interface of Robonaut-2 targets telepresence.

19.4.7 Telerobotics with a Rover

Rovers have also been managed under the concept of telerobotics and supervisory control, whereby the operator plays a crucial role. There are three key points compared with space telemanipulation

- (1) The workplace is far from Earth.
- (2) The rover operates in an unknown environment.
- (3) The rover collects explorative information and sends it to Earth.

It is unreasonable to send continuous low-level commands to a rover on Mars, as the communication time delay can be several minutes. This increases the value of autonomous capabilities. Moreover, it is impossible to provide a preliminary remote environment model, meaning more advanced supervisory control is required. The key technology is simultaneous localization and mapping (SLAM), whereby localization and mapping is provided using a laser range-finder or stereo camera.

The main purpose of the rover is exploration, which requires high-level decision making. The rover supplies useful information to the scientists involved in the project by satisfying their requirements. The exploration of Mars by rovers started with *Sojourner*, which was followed by *Spirit*, then *Opportunity* (which remained operational for an unexpectedly long time), and then *Curiosity*. *Curiosity* is significantly larger than its predecessors and can travel greater distances, showing that the level of autonomy is rapidly improving.

References

1. D. L. Akin, M. L. Minsky, E. D. Thiel and C. R. Curtzman, "Space Applications of Automation, Robotics and Machine Intelligence Systems (ARAMIS) phase II," NASA-CR-3734 - 3736, 1983.
2. P. Laryssa, E. Lindsay, O. Layi, O. Marius, K. Nara, L. Aris, T. Ed, "International Space Station Robotics: A Comparative Study of ERA, JEMRMS and MSS", Proc. 7th ESA Workshop on Advanced Space Technologies for Robotics and Automation, ASTRA 2002, ESTEC, Noordwijk, The Netherlands, Nov. 2002.
3. Hirzinger, G., Brunner, B., Dietrich, J., Heindl, J., "Sensor-based Space Robotics-ROTEX and its Telerobotic Features," IEEE Trans. on Robotics and Automation, Vol. 9, No. 5, pp.649-663, 1993.
4. Preusche, C., Reintsema, D., Landzettel, K., Hirzinger, G., "Robotics Component Verification on ISS ROKVISS - Preliminary Results for Telepresence," Proc. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4595 - 4601, 9-15 Oct. 2006
5. M. Oda et al, "ETS-VII, Space Robot In-Orbit Experiment Satellite," Proc. 1996 IEEE Int. Conf. on Robotics and Automation, pp.739-744, 1996.
6. K. Yoshida, "Engineering Test Satellite VII Flight Experiments For Space Robot Dynamics and Control: Theories on Laboratory Test Beds Ten Years Ago, Now in Orbit," International Journal of Robotics Research, Vol. 22, No. 5, pp. 321-335, 2003.
7. Robert B. Friend, "Orbital Express Program Summary And Mission Overview," Sensors and Systems for Space Applications II, Proc. of SPIE Vol. 6958, 695803, (2008)
8. Diftler, M.A., Mehling, J.S., Abdallah, M.E., Radford, N.A., Bridgwater, L.B., Sanders, A.M., Askew, R.S., Linn, D.M., Yamokoski, J.D., Permenter, F.A., Hargrave, B.K., Piatt, R., Savely, R.T., Ambrose, R.O., "Robonaut 2 - The first humanoid robot in space," Proc. 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 2178-2183, 9-13 May 2011
9. "Lunokhod 1 Mobile Lunar Laboratory, USSR" published by the Joint Publications Research Service (JPRS identification number 54525.), US Department of Commerce, a translation of a Russian language monograph: Peredvizhnaya Laboratoriya na Lune Lunokhod-1, responsible Editor Academician A. P. Vinogradov, Moscow, 4 June 1971, 128 pages.
10. Mishkin, A., *Sojourner: An Insider's View of the Mars Pathfinder Mission*, Berkley Books, 2004, ISBN 0-425-19199-0
11. Maimone M., Yang Cheng Y., Matthies L., "Two years of Visual Odometry on the Mars Exploration Rovers," Journal of Field Robotics, Vol. 24, No. 3, pp. 169186, 2007.
12. Mars Exploration Rovers Mission Homepage; <http://marsrover.nasa.gov/home/> (as of August 2012)
13. Mars Science Laboratory Curiosity Rover Homepage; <http://mars.jpl.nasa.gov/msl/> (as of August 2012)
14. Kuninaka, H. and Kawaguchi, J., "Lessons learned from round trip of Hayabusa asteroid explorer in deep space," Proc. 2011 IEEE Aerospace Conference, pp. 1-8, 5-12 March 2011
15. Hajime Yano, T. Kubota, H. Miyamoto, T. Okada, D. Scheeres, Y. Takagi, K. Yoshida, M. Abe, S. Abe, O. Barnouin-Jha, A. Fujiwara, S. Hasegawa, T. Hashimoto, M. Ishiguro, M. Kato, J. Kawaguchi, T. Mukai, J. Saito, S. Sasaki and M. Yoshikawa, "Touchdown of the Hayabusa Spacecraft at the Muses Sea on Itokawa," Science, Vol. 312, no. 5778, pp. 1350-1353, 2 June 2006
16. K. Yoshida, T. Kubota, S. Sawai, A. Fujiwara, M. Uo, "MUSES-C Touch-down Simulation on the Ground," AAS/AIAA Space Flight Mechanics Meeting, Paper AAS 01-135, Santa Barbara, California, pp. 1-10, February 2001.
17. T. Nakamura, T. Noguchi, M. Tanaka, M. E. Zolensky, M. Kimura, A. Tsuchiyama, A. Nakato, T. Ogami, H. Ishida, M. Uesugi, T. Yada, K. Shirai, A. Fujimura, R. Okazaki, S. A. Sandford, Y. Ishibashi, M. Abe, T. Okada, M. Ueno, T. Mukai, M. Yoshikawa, J. Kawaguchi, "Itokawa Dust Particles: A Direct Link Between S-Type Asteroids and Ordinary Chondrites," Science, Vol. 333, no. 6046, pp. 1113-1116, 26 August 2011
18. W. Book, "Structural flexibility of motion systems in the space environment," IEEE Transactions on Robotics and Automation, vol. 9, no. 5, pp. 524-530, 1993.
19. S. Cetinkunt and W. Book, "Transfer Functions of Flexible Beams and Implications of Flexibility on Controller Performance," in Teleoperation and Robotics in Space (S. B. Skaar and C. F. Ruoff, eds.), pp. 291-313, AIAA, 1994.
20. S. Ulrich, J. Z. Sasiadek, and I. Barkana, "Modeling and Direct Adaptive Control of a Flexible-Joint Manipulator," Journal of Guidance, Control, and Dynamics, vol. 35, pp. 25-39, Jan. 2012.
21. W. J. Book, O. Maizza-Neto, and D. E. Whitney, "Feedback Control of Two Beam, Two Joint Systems With Distributed Flexibility," Journal of Dynamic Systems, Measurement, and Control, vol. 97, no. 4, p. 424, 1975.
22. Y. Chen and L. Meirovitch, "Control of a flexible space robot executing a docking maneuver," Journal of Guidance, Control, and Dynamics, vol. 18, pp. 756-766, July 1995.
23. L. Meirovitch and T. Stemple, "Hybrid equations of motion for flexible multibody systems using quasicordinates," Journal of Guidance, Control, and Dynamics, vol. 18, pp. 678-688, July 1995.
24. R. Masoudi and M. Mahzoon, "Maneuvering and Vibrations Control of a Free-Floating Space Robot with Flexible Arms," Journal of Dynamic Systems, Measurement, and Control, vol. 133, no. 5, 2011.

25. S. H. Lee and W. J. Book, "Robot Vibration Control Using Inertial Damping Forces," in VIII CISM-IFTToMM Symposium on the Theory and Practice of Robots and Manipulators (Ro. Man. Sy. '90), (Cracow, Poland), pp. 252–259, 1990.
26. R. Lindberg, R. Longman, and M. Zedd, "Kinematic Dynamic Properties of an Elbow Manipulator Mounted on a Satellite," in Space Robotics: Dynamics and Control (Y. Xu and T. Kanade, eds.), pp. 126, Boston MA: Kluwer Academic Pubs., 1993.
27. D. Nenchev, Y. Umetani, and K. Yoshida, "Analysis of a redundant free-flying spacecraft/manipulator system," IEEE Transactions on Robotics and Automation, vol. 8, no. 1, pp. 16, 1992.
28. R. Featherstone, Rigid Body Dynamics Algorithms. Boston, MA: Springer US, 2008.
29. D. Nenchev, K. Yoshida, P. Vichitkulsawat, and M. Uchiyama, "Reaction null-space control of flexible structure mounted manipulator systems," IEEE Transactions on Robotics and Automation, vol. 15, no. 6, pp. 1011–1023, 1999.
30. Z. Vafa and S. Dubowsky, "On the dynamics of manipulators in space using the virtual manipulator approach," in Proceedings. 1987 IEEE International Conference on Robotics and Automation, vol. 4, pp. 579–585, Institute of Electrical and Electronics Engineers, 1987.
31. Y. Umetani and K. Yoshida, "Resolved motion rate control of space manipulators with generalized Jacobian matrix," IEEE Transactions on Robotics and Automation, vol. 5, pp. 303–314, June 1989.
32. Y. Masutani, F. Miyazaki, and S. Arimoto, "Modeling and sensory feedback control for space manipulators," in NASA Conf. Space Telerobotics, 1989.
33. K. Yoshida, "Impact dynamics representation and control with Extended Inversed Inertia Tensor for space manipulators," in Robotics Research: The 6th International Symposium (R. Paul and T. Kanade, eds.), pp. 453–463, The International Foundation for Robotics Research, 1994.
34. K. Yoshida, R. Kurazume, and Y. Umetani, "Dual arm coordination in space free-flying robot," in Proceedings. 1991 IEEE International Conference on Robotics and Automation, no. April, pp. 2516–2521, IEEE Comput. Soc. Press, 1991.
35. Y. Yokokohji, T. Toyoshima, and T. Yoshikawa, "Efficient computational algorithms for trajectory control of free-flying space robots with multiple arms," IEEE Transactions on Robotics and Automation, vol. 9, no. 5, pp. 571–580, 1993.
36. E. Papadopoulos and S. A. a. Moosavian, "Dynamics and control of space free-flyers with multiple manipulators," Advanced Robotics, vol. 9, pp. 603–624, Jan. 1994.
37. S. K. Agrawal and S. Shirumalla, "Planning motions of a dual-arm free-floating manipulator keeping the base inertially fixed," Mechanism and Machine Theory, vol. 30, pp. 5970, Jan. 1995.
38. B. Siciliano and O. Khatib, eds., Springer Handbook of Robotics. No. D, Berlin, Heidelberg: Springer Verlag, 2008.
39. Y. Nakamura and R. Mukherjee, "Nonholonomic path planning of space robots via a bidirectional approach," IEEE Transactions on Robotics and Automation, vol. 7, no. 4, pp. 500–514, 1991.
40. W. R. Doggett, W. C. Messner, and J.-N. Juang, "Global minimization of robot base reaction forces during point to point moves," in AIAA Guidance, Navigation and Control Conference, (San Diego, California), pp. AIAA963–896, 1996.
41. K. Yoshida, H. Nakanishi, H. Ueno, N. Inaba, T. Nishimaki, and M. Oda, "Dynamics, control and impedance matching for robotic capture of a non-cooperative satellite," Advanced Robotics, vol. 18, pp. 175–198, Jan. 2004.
42. N. Hogan, "Impedance Control: An Approach to Manipulation: Part I-Theory," Journal of Dynamic Systems, Measurement, and Control, vol. 107, no. 1, p. 1, 1985.
43. S. A. A. Moosavian, E. Papadopoulos, and R. Rastegari, "Multiple Impedance Control for Space Free-Flying Robots," Journal of Guidance, Control, and Dynamics, vol. 28, pp. 939–947, Sept. 2005.
44. D. Nenchev and K. Yoshida, "Impact analysis and post-impact motion control issues of a free-floating Space robot subject to a force impulse," IEEE Transactions on Robotics and Automation, vol. 15, pp. 548–557, June 1999.
45. D. Dimitrov and K. Yoshida, "Utilization of the bias momentum approach for capturing a tumbling satellite," 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), vol. 4, pp. 3333–3338, 2004.
46. Muir, P., and Neuman, C., "Kinematic modeling of wheeled mobile robots," Technical Report, CMU-RI-TR-86-12, Robotics Institute, Carnegie Mellon University, June, 1986.
47. Alexander, J. C., and Maddocks, J.H., "On the kinematics of wheeled mobile robot," International Journal of Robotics Research, Vol. 8, No. 5, 1989, pp.15–27.
48. Fierro, R., and Lewis, F. L., "Control of a Nonholonomic Mobile Robot: Backstepping Kinematics into Dynamics," Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp.3805–3810.
49. Tarokh, M., and McDermott, G. J., "Kinematics Modeling and Analyses of Articulated Rovers," IEEE Transactions on Robotics, Vol. 21, No. 4, 2005, pp. 539–553.
50. Chakraborty, N., and Ghosal, A., "Kinematics of Wheeled Mobile Robot on Uneven Terrain," Mechanism and Machine Theory, Vol. 39, Issue 12, 2004, pp. 1273–1287
51. Grand, C., Benamar, Faiz., and Plumet, F., "Motion Kinematics Analysis of Wheeled-Legged Rover over 3D Surface with Posture Adaptation," Mechanism and Machine Theory, Vol. 45, Issue 3, 2010, pp. 477–495.
52. Jain, A., Balaram, J., Cameron, J., Guineau, J., Lim, C., Pomerantz, M., and Sohl, G., "Recent Developments in the ROAMS Planetary Rover Simulation Environment," Proceedings of the 2004 IEEE Aerospace Conference, Big Sky, MT, 2004, pp. 861–876.
53. Bauer, R., Leung, W., and Barfoot, T., "Development of a Dynamic Simulation Tool for the Exomars Rover," Proceedings of the 8th International. Symposium on Artificial Intelligence, Robotics and Automation in Space, Munich, Germany, 2005.
54. Gibbesch, A., and Schäfer, B., "Multibody system modelling and simulation of planetary rover mobility on soft terrain," Proceedings of the 8th International. Symposium on Artificial Intelligence, Robotics and Automation in Space, Munich, Germany, 2005.
55. Krenn, R., Gibbesch, A., and Hirzinger, G., "Contact Dynamics Simulation of Rover Locomotion," Proceedings of the 9th International. Symposium on Artificial Intelligence, Robotics and Automation in Space, Los Angeles, CA, 2007.
56. Bickler, D., "Articulated suspension system," US Patent 4840394, 1988.
57. Volpe, R., Balaram, J., Ohm, T., and Ivlev, R., "Rocky 7: A Next Generation Mars Rover Prototype," Advanced Robotics, Vol. 11, No. 4, 1997, pp. 341–358.
58. Harrington, B. D., and Voorhees, C., "The Challenges of Designing the Rocker-Bogie Suspension for the Mars Exploration Rover," NASA Technical Documents, 20040084284, Nov., 2005.
59. Iagnemma, K., and Dubowsky, S., "Mobile Robots in Rough Terrain: Estimation, Motion Planning, and Control with application to Planetary Rovers," Springer Tracts in Advanced Robotics (STAR) Series, Vol. 12, Springer, 2004.

60. Ishigami, G., Nagatani, K., and Yoshida, K., "Slope Traversal Controls for Planetary Exploration Rover on Sandy Terrain," *Journal of Field Robotics*, Vol. 26, Issue 3, pp. 264–286, 2009.
61. Coulter, R. Craig, "Implementation of the Pure Pursuit Path Tracking Algorithm," CMU-RI-TR-92-01, 1992.
62. Helmick, D. M., Roumeliotis, S. I., Cheng, Y., Clouse, D. S., Bajracharya, M., and Matthies, L. H., "Slip Compensated Path Following for Planetary Exploration Rovers," *Advanced Robotics*, Vol. 20, pp. 1257–1280, 2006.
63. Ishigami, G., Miwa, A., Nagatani, K., and Yoshida, K., "Terramechanics-Based Model for Steering Maneuver of Planetary Exploration Rovers on Loose Soil," *Journal of Field Robotics*, Vol. 24, Issue 3, 2007, pp. 233–250.
64. "Mars Science Laboratory Fact Sheet". NASA/JPL. Retrieved Oct. 16th, 2012.
65. Bekker, M. G., "Theory of Land Locomotion," Ann Arbor, MI, University of Michigan Press, 1956.
66. Bekker, M. G., "Introduction to Terrain-Vehicle Systems," Ann Arbor, MI, University of Michigan Press, 1969.
67. Wong, J.Y., "Theory of Ground Vehicles," 4th edn., Wiley, New York, 2008.
68. Wong, J.Y., and Reece A.R., "Prediction of Rigid Wheel Performance based on the Analysis of Soil-Wheel Stresses Part I, Performance of Driven Rigid Wheels," *Journal of Terramechanics*, Vol. 4, No. 1, 1967, pp. 81–98.
69. Wong, J.Y., and Reece A.R., "Prediction of Rigid Wheel Performance based on the Analysis of Soil-Wheel Stresses Part II, Performance of Towed Rigid Wheels," *Journal of Terramechanics*, Vol. 4, No. 2, 1967, pp. 7–25.
70. Schmid, I. C., "Interaction of Vehicle and Terrain Results from 10 Years Research at IKK," *Journal of Terramechanics*, Vol. 32, No. 1, 1995, pp. 3–25.
71. Ding, L., Deng, Z., Gao, H., Nagatani, K., and Yoshida, K., "Planetary rovers' wheel-soil interaction mechanics: new challenges and applications for wheeled mobile robots," *Intelligent Service Robotics*, Vol. 4 Issue 1, Springer-Verlag New York, December, 2010, pp. 17–38.
72. Nakashima, H., Fujii, H., Oida, A., Momozu, M., Kawase, Y., Kanamori, H., Aoki, S., and Yokoyama, T., "Parametric analysis of lugged wheel performance for a lunar microrover by means of DEM," *Journal of Terramechanics*, Vol. 44, 2007, pp. 153–162.
73. Nakashima, H., Fujii, H., Oida, A., Momozu, M., Kanamori, H., Aoki, S., Yokoyama, T., Shimizu, H., Miyasaka, J., and Ohdoi, K., "Discrete element method analysis of single wheel performance for a small lunar rover on sloped terrain," *Journal of Terramechanics*, Vol. 47, 2010, pp. 307–321, 2010.
74. Li, W., Huang, Y., Cui, Y., Dong, S., and Wang, J., "Trafficability analysis of lunar mare terrain by means of the discrete element method for wheeled rover locomotion," *Journal of Terramechanics*, Vol. 47, 2010, pp. 161–172.
75. Wakabayashi, S., Sato, H., and Nishida, S., "Design and mobility evaluation of tracked lunar vehicle," *Journal of Terramechanics*, Vol. 46, Issue, 3, 2009, pp. 105–114.
76. Patel, N., Slade, R., and Clemmet, Jim., "The ExoMars rover locomotion subsystem," *Journal of Terramechanics*, Vol. 47, 2010, pp. 227–242.
77. Lindemann, R., Bickler, D., Harrington, B., Ortiz, G., and Voorhees, C., "Mars Exploration Rover Mobility Development," *IEEE Robotics and Automation Magazine*, Vol.13, Issue 2, June, 2006, pp. 19–26.
78. Ishigami, G., Miwa, A., Nagatani, K., and Yoshida, K., "Terramechanics-Based Analysis on Slope Traversability for a Planetary Exploration Rover," *Proceedings of the 25th International Symposium on Space Technology and Science*, 2006, pp. 1025–1030.
79. Michaud, S., Richter, L., Thueer, T., Gibbesch, A., Huelsing, T., Schmitz, N., Weiss, S., Krebs, A., Patel, N., Joudrier, L., Siegwart, R., Schäfer, B., and Ellery, A., "Rover Chassis Evaluation and Design Optimisation using the RCET," *Proceedings of the 9th ESA Workshop on ASTRA*, 2006.
80. Iagnemma, K., Senatore, C., Trease, B., Arvidson, R., Shaw, A., Zhou, F., Van Dyke, L, and Lindemann, R., "Terramechanics Modeling of Mars Surface Exploration Rovers for Simulation and Parameter Estimation," *Proceedings of the ASME International Design Engineering Technical Conference*, 2011.
81. Onafeko, O., and Reece A.R., "Soil Stresses and Deformations beneath Rigid Wheels," *Journal of Terramechanics*, Vol. 4, No. 1, 1967, pp. 59–80.
82. Janosi, Z., and Hanamoto, B., "Analytical determination of drawbar pull as a function of slip for tracked vehicle in deformable soils," *Proceedings of the 1st ISTVS Conference*, Torino, Turin, Italy, 1961, pp. 707–726.
83. Iagnemma, K., "A Laboratory Single Wheel Testbed for studying planetary rover wheel-terrain interaction," *Massachusetts Institute of Technology Technical Report*, 01-05-05, 2005.
84. Nagatani, K., Ikeda, A., Sato, K., and Yoshida, K., "Accurate Estimation of Drawbar Pull of Wheeled Mobile Robots Traversing Sandy Terrain Using Built-in Force Sensor Array Wheel," *Proceedings of the 2009 IEEE/RSJ International Conference on Robots and Systems*, St. Louis, MO, 2009, pp. 2373–2378
85. Meirion-Griffith, G., and Spenko, M., "A Modified Pressure-sinkage Model for Small, Rigid Wheels on Deformable Terrains," *Journal of Terramechanics*, Vol. 48, Issue 2, 2011, pp. 149–155.
86. Senatore, C., and Iagnemma, K., "Direct shear behaviour of dry, granular soils for low normal stress with application to lightweight robotic vehicle modeling," *Proceedings of the 17th ISTVS Conference*, Blacksburg, VA, 2011.
87. Matijevic J., Crisp J., Bickler D., Baner R., Cooper B., Eisen H., Gensler J., Haldemann A., Hartman F., Jewett K., Matthies L., Laubach S., Mishkin A., Morrison J., Nguyen T., Sirota A., Stone H., Stride S., Sword L., Tarsala J., Thompson A., Wallace M., Welch R., Wellman E., Wilcox B., Foerguson D., Jenkins P., Kolecki J., Landis G., and Wilt D., "Characterization of Martian surface deposits by the Mars pathfinder rover, Sojourner," *Science*, Vol. 278, No. 5, 1997, pp. 1765–1768.
88. Iagnemma K., Kang S., Shibly H., and Dubowsky S., "Online terrain parameter estimation for wheeled mobile robots with application to planetary rovers," *IEEE Transactions on Robotics*, Vol. 20, No. 5, 2004, pp. 921–927.
89. Hutangkabodee S., Zweiri Y., Seneviratne L., and Althoefer K., "Soil parameter identification for wheel-terrain interaction dynamics and traversability prediction," *International Journal of Automation and Computing*, Vol. 3, No. 3, 2006, pp. 244–251.
90. Helmick D., Angelova A., Matthies L., Brooks C., Halatci I., Dubowsky S., and Iagnemma K., "Experimental results from a terrain adaptive navigation system for planetary rovers," *Proceedings of 9th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, i-SAIRAS, Hollywood, CA, 2008, No. 110.
91. Ishigami, G., Kewlani, G., and Iagnemma, K., "A Statistical Approach to Mobility Prediction for Planetary Surface Exploration Rovers in Uncertain Terrain," *IEEE Robotics and Automation Magazine*, Vol.16, Issue 4, December, 2009, pp. 61–70.
92. Matthies, L., "Stereo Vision for Planetary Rovers: Stochastic Modeling to Near Real-Time Implementation," *International Journal of Computer Vision*, Vol. 8, No. 1, 1992 pp. 71–91.

93. Golberg, S., Maimone, M., and Matthies, L., "Stereo Vision and Rover Navigation Software for Planetary Exploration," Proceedings of the 2002 IEEE Aerospace Conference, Big Sky, MT, 2002, pp. 5-2025-5-2036.
94. Maimone, M., Johnson, A., Cheng, Y., Willson, R., and Matthies, L., "Autonomous Navigation Results from the Mars Exploration Rover (MER) Mission," 9th International Symposium on Experimental Robotics, Singapore, 2004.
95. Li, R., Squyres, S., Arvidson, R., Archinal, B., Bell, J., Cheng, Y., Crumpler, L., Marais, D., Di, K., Ely, T., Golombek, M., Graat, E., Grant, J., Guinn, J., Johnson, A., Greeley, R., Kirk, R., Maimone, M., Matthies, L., Malin, M., Parker, T., Sims, M., Soderblom, L., Thompson, S., Wang, J., Whelley, P., and Xu, F., "Initial results of rover localization and topographic mapping for the 2003 Mars exploration rover mission," Photogrammetric Engineering & Remote Sensing (Special Issue on Mapping Mars), Vol. 71, No. 10, 2005, pp. 1129-1142.
96. Matthies, L., Maimone, M., Johnson, A., Cheng, Y., Willson, R., Villalpando, C., Goldberg, S., Huertas, A., Stein, A., and Angelova, A., "Computer Vision on Mars," International Journal of Computer Vision, Vol. 75, No. 1, 2007, pp. 67-92.
97. Wulf, O., and Wagner, B., "Fast 3D Scanning Methods for Laser Measurement Systems," Proceedings of the International Conference on Control Systems and Computer Science, Bucharest, Romania, 2003, pp. 312-317.
98. Thrun, S., Thayer, S., Whittaker, W., Baker, C., Burgard, W., Ferguson, D., Hahnel, D., Montemerlo, M., Morris, A., Omohundro, Z., Reverte, C., and Whittaker, W., "Autonomous Exploration and Mapping of Abandoned Mines," IEEE Robotics and Automation Magazine, Vol.11, No.4, 2004, pp. 79-91.
99. Buehler, M., Iagnemma, K., and Singh, S., (eds.) "The 2005 DARPA Grand Challenge: The Great Robot Race," Springer Tracts in Advanced Robotics (STAR) Series, Vol. 36, Springer-Verlag Berlin Heidelberg, 2005.
100. Buehler, M., Iagnemma, K., and Singh, S., (eds.) "The DARPA Urban Challenge: Autonomous Vehicles in City Traffic," Springer Tracts in Advanced Robotics (STAR) Series, Vol. 56, Springer-Verlag Berlin Heidelberg 2009.
101. "DragonEye 3D Flash LIDAR Space Camera," <http://www.advancedscientificconcepts.com/products/dragoneye.html>, Retrieved Oct. 16th, 2012.
102. Luu, T., Ruel, S., and Berube, A., "TriDAR Test Results Onboard Final Shuttle Mission, Applications For Future Of Non-Cooperative Autonomous Rendezvous & Docking," Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space 2012.
103. Vaskevicius, N., Birk, A., Pathak, K., and Schwertfeger, S., "Efficient Representation in 3D Environment Modeling for Planetary Robotic Exploration", Advanced Robotics, vol. 24, no. 8-9, pp. 1169-1197, 2010.
104. Dong, H., and Barfoot, T., "Lighting-Invariant Visual Odometry using Lidar Intensity Imagery and Pose Interpolation," Proceedings of the 8th International Conference on Field and Service Robotics, July, 2012.
105. Bakambu, J., Nimelman, M., Tripp, J., and Kujele, A., "Compact fast scanning lidar for planetary rover navigation," Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space 2012.
106. Angelova, A., Matthies, L., Helmick, D., and Perona, P., "Learning slip behavior using automatic mechanical supervision," Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Rome, 2007, pp. 1741-1748.
107. Angelova, A., Matthies, L., Helmick, D., Sibley, G., and Perona, P., "Learning to predict slip for ground robots," Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, FL, 2006, pp. 3324-3331.
108. Li, R., Di, K., Matthies, L., Arvidson, R., Folkner, W., Archinal, B., "Rover Localization and Landing Site Mapping Technology for the 2003 Mars Exploration Rover Mission," Journal of Photogrammetric Engineering and Remote Sensing, Vol. 70, No. 1, 2004, pp. 77-90.
109. Olson, C., Matthies, L., and Schoppers, M., "Rover Navigation using Stereo Ego-motion," Robotics and Autonomous Systems, Vol. 43, No. 4, 2003, pp. 215-229.
110. Maimone, M., Cheng, Y., and Matthies, L., "Two years of visual odometry on the Mars Exploration Rovers," Journal of Field Robotics, Vol. 24, Issue 3, 2007, pp. 169-186.
111. Stentz, A., "Optimal and efficient path planning for partially-known environments," Proceedings of the 1994 IEEE International Conference on Robotics and Automation, San Diego, CA, 1994, pp. 3310-3317.
112. Barraquand, J., Langlois, B., and Latombe, J., "Numerical potential field techniques for robot path planning," IEEE Transactions on Systems, Man and Cybernetics, Vol. 22, No. 2, 1992, pp. 224-241.
113. Kavraki, L., Svestka, P., Latombe, J., and Overmars, M., "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," IEEE Transactions on Robotics and Automation, Vol. 12, No. 4, 1996, pp. 566-580.
114. Cheng, P., Shen, Z., and LaValle, S., "RRT-based trajectory design for autonomous automobiles and spacecraft," Archives of Control Sciences, Vol. 11, No. 3-4, 2001, pp. 167-194.
115. Donald, B., Xavier, P., Canny, J., and Reif, J., "Kinodynamic motion planning," Journal of the Association for Computing Machinery, Vol. 40, No. 5, 1993, pp. 1048-1066.
116. Urmson, C., and Simmons, R., "Approaches for heuristically biasing RRT growth," Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, 2003, pp. 1178-1183.
117. Kuwata, Y., Teo, J., Karaman, S., Fiore, G., Frazzoli, E., and How, J., "Real-time Motion Planning with Applications to Autonomous Urban Driving," IEEE Transaction on Control Systems Technology, Vol. 17, No. 5, 2009, pp. 1105-1118.
118. Howard, T., and Kelly, A., "Trajectory and Spline Generation for All-Wheel Steering Mobile Robots," Proceedings of the 2006 IEEE International Conference on Intelligent Robots and Systems, Beijing, China, 2006, pp. 4827-4832.
119. Howard, A., Seraji, H., and Tunstel, E., "A Rule-Based Fuzzy Traversability Index for Mobile Robot Navigation," Proceedings of the 2001 IEEE Conference on Robotics and Automation, Seoul, Korea, 2001, pp. 3067-3071.
120. Singh, S., Simmons, R., Smith, T., Stentz, A., Verma, V., Yahja, A., and Schwehr, K., "Recent Progress in Local and Global Traversability for Planetary Rovers," Proceedings of the 2000 IEEE Conference on Robotics and Automation, San Francisco, CA, 2000, pp. 1194-1200.
121. Ishigami, G., Nagatani, K., and Yoshida, K., "Path Planning and Evaluation for Planetary Rovers Based on Dynamic Mobility Index," Proceedings of the IEEE International Conference on Robots and Systems, San Francisco, CA, 2011, pp. 601-606 .
122. Biesiadecki, J., and Maimone, M., "The Mars Exploration Rover surface mobility flight software: Driving ambition," Proceedings of the 2006 IEEE Aerospace Conference, Big Sky, MT, 2006.
123. Carsten, J., Rankin, A., Ferguson, D., Stentz, A., "Global Path Planning on Board the Mars Exploration Rovers," Proceedings of the 2007 IEEE Aerospace Conference, 2007, pp. 1-11.
124. Carsten, J., Rankin, A., Ferguson, D., Stentz, A., "Global planning on the Mars Exploration Rovers: Software integration

- and surface testing,” *Journal of Field Robotics*, Vol. 26, Issue 4, 2008, pp. 337–357.
125. Kelly, A., Stentz, A., Amidi, O., Bode, M., Bradley, D., Diaz-Calderon, A., Happold, M., Herman, H., Mandelbaum, R., Pilarski, T., Rander, P., Thayer, S., Vallidis, N., and Warner, R., “Toward reliable off road autonomous vehicles operating in challenging environments,” *International Journal of Robotics Research*, Vol. 25, No. 5-6, 2006, pp.449–483.
 126. Ishigami, G., Otsuki, M., and Kubota, T., “Path Planning and Navigation Framework for a Planetary Exploration Rover using a Laser Range Finder,” *Proceedings of the 8th International Conference on Field and Service Robotics*, July, 2012.
 127. W. R. Ferrel, “Remote Manipulation with Transmission Delay,” *IEEE Trans. on Human Factors in Electronics*, HFE-6, No. 1, pp. 24–32, 1965.
 128. M. V. Noyes and T. B. Sheridan, “A Novel Predictor for Telemanipulation through a Time Delay,” *Proc. of Annual Conf. on Manual Control*.
 129. G. Hirzinger, B. Brunner, J. Dictrich and J. Heindl, “Sensor-Based Space Robotics-ROTEX and Its Telerobotic Features,” *IEEE Trans. on Robotics and Automation*, Vol. 9, No. 5, pp. 649–663, 1993.
 130. M. Oda and T. Doi, “Teleoperation System of ETS-VII Robot Experiment Satellite,” *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 1644–1650, 1997.
 131. R. J. Anderson and M. W. Spong, “Bilateral Control of Teleoperators with Time Delay,” *IEEE Trans. on Automatic Control*, Vol. 34, No. 5, pp. 494–501, 1989.
 132. W. K. Yoon, T. Goshozono, H. Kawabe, M. Kinami, Y. Tsumaki, M. Uchiyama, M. Oda, T. Doi, “Model-Based Teleoperation of ETS-VII Manipulator,” *IEEE Trans. on Robotics and Automation*, Vol. 20, No. 3, pp. 602–612, 2004.
 133. T. Imaida, Y. Yokokohji, T. Doi, M. Oda, T. Yoshikawa, “Ground-space bilateral teleoperation of ETS-VII robot arm by direct bilateral coupling under 7-s time delay condition,” *IEEE Trans. on Robotics and Automation*, Vol. 20, No. 3, pp. 499–511, 2004.
 134. G. Hirzinger, K. Landzettel, D. Reinsema, C. Preusche, A. Albu-Schaffer, B. Rebele, M. Turk, “ROKVISS-Robotics Component Verification on ISS,” *i-SAIRAS2005*, 2005.
 135. W. R. Ferrel and T. B. Sheridan, “Supervisory Control of Remote Manipulation,” *IEEE Spectrum*, Vol. 4, No. 10, pp. 81–88, 1967.
 136. T. B. Sheridan, “Telerobotics, Automation, and Human Supervisory Control,” *The MIT Press*.
 137. A. K. Bejczy, W. S. Kim, S. C. Venema, “The phantom robot: predictive displays for teleoperation with time delay,” *1990 IEEE Int. Conf. on Robotics and Automation*, pp. 546–551, 1990.
 138. E. Freund and J. Rossmann, “Projective virtual reality: bridging the gap between virtual reality and robotics,” *IEEE Trans. on Robotics and Automation*, Vol. 15, No. 3, pp. 411–422, 1999.

Richard Lowe, Dan Kent, Paul Coutinho and Kevin Halsall

This chapter provides an overview of systems, processes and techniques used within the mission ground segment.

As introduced in [Chap. 2](#), the ground segment comprises those elements of the mission system that are used to control the spacecraft and its payload, and to process the data returned from it. Note that launch infrastructure is not normally considered to be part of the ground segment.

The ‘technology’ focus of the ground segment is mainly on IT-based systems and communications, but these tools are primarily a ‘means to an end’. Operators of the mission must have a strong, broad, and deep understanding of the design of the spacecraft itself—particularly with respect to the on-board software, interconnectivity of subsystems and their relative physical configuration. This technical knowledge is combined with a strong ‘human factors’ element and a need for robust processes and procedures to be designed and followed by the engineers and managers who fly the mission.

Activities can be divided into two general domains: control and monitoring of the satellite platform (implemented through the flight operations segment, sometimes also referred to as the ground control segment); and operation and exploitation of the payload (implemented through the payload data ground segment, or ground mission segment). For both satellite platform and payload, it is normal to make extensive use of standard IT hardware, but with software that is developed specifically for spacecraft operations and data processing.

Radio-frequency and optical technologies are also used for communications links and tracking of spacecraft, and these are addressed in [Chap. 14](#).

It should be noted that, unlike other areas of space engineering, the approach to ground segment software and

operational practices for a mission is as much a methodology choice as a technological one.

This chapter based upon the approach adopted by the European Space Agency.

20.1 Flight Operations Segment

This section provides an overview of the operation of the spacecraft platform. Note that the operation of the spacecraft *payload* may require additional distinct systems and processes of its own—e.g. for the management of data channels on a communications payload, or for the generation of a navigation signal in the case of systems like GPS or Galileo. Nevertheless, in most cases the physical hardware of the payload is managed through the same systems, and by the same team, as the main spacecraft platform.

The goal of the flight operations segment is to give the spacecraft operator the tools required to reliably and efficiently control the spacecraft in order to support the mission’s aims. Successful control of the platform is a necessary (though not normally sufficient) condition for completing the mission. As an example, it is instructive to consider the case of a geostationary communications satellite. The system will be used to ensure that the satellite remains in the correct orbit and in the correct orientation, that sufficient power is always available to the payload, and that the operational lifetime of the satellite is maximized. These goals ensure that the payload is supported in doing its job—however, they are independent of the payload’s own complex task, which is to receive and retransmit data channels for terrestrial users (e.g. television broadcast, mobile phones, Internet traffic, etc.).

In other mission types, the operation of the spacecraft platform and the operation of the payload may be tightly coupled. As an example, Earth observation satellites may require the whole satellite to reorientate itself for each

R. Lowe (✉) · D. Kent · P. Coutinho · K. Halsall
Telespazio VEGA UK Ltd. Luton, England
e-mail: Richard.Lowe@vegaspaces.com

image acquisition. In such cases, the distinction between operation of the payload and the platform is far weaker.

In either case, the manner in which operators go about controlling the spacecraft is as important as the underlying technology. To draw a terrestrial parallel, a mission may be thought of in similar terms to a remotely operated aircraft. The aircraft itself embodies the majority of the aeronautical engineering technology. The operator's control system must be designed in order to provide the best possible interface for flying the aircraft. The operator must adopt suitable practices for minimizing risk of human error.

20.1.1 Flight Operations Procedures

This section introduces the concept of the flight operations procedure (FOP). Note that this abbreviation FOP is sometimes also used to refer to the flight operations plan, which is a document that contains all the spacecraft procedures.

Spacecraft procedures are designed to prevent critical errors from being made when commanding the spacecraft. They describe, in step-by-step detail, the process of commanding the spacecraft (and ground systems) in order to achieve a specific objective. It is common practice for spacecraft operators to use procedures as the basis of every command that they send. The implications of sending the wrong command (or omitting a required command) can be very serious indeed.

A procedure will normally include the following items of information

- A description of the task that the procedure is designed to carry out (e.g. switch on payload).
- A description of start conditions, describing the state that the spacecraft and/or the unit that is being commanded must be in, before beginning the procedure.
- A description of end conditions, describing the state that the spacecraft and/or the unit that is being commanded will be in, after successfully completing the procedure.
- At each step of the procedure
 - Telemetry to check and values to expect.
 - Telecommands to send and parameters to set.
 - Timing and conditions for commanding, telemetry and on-board events ('IF/THEN', 'WAIT' steps etc.).
- References to other procedures that may be called on to perform subsidiary tasks (e.g. 'Prepare for Payload switch-on').
- References to other documents that contain detailed operational information, for example memory addresses to be set in telecommand parameters. In short, the procedure must give as much relevant detail as possible to the spacecraft controller or operations engineer.

For infrequently used procedures, literally *years* may elapse between the writing of the procedure and its use to command the spacecraft. Thus, it is very important to record as much knowledge as possible as a reminder for when that day arrives.

Most procedures will be developed in the period up to three years prior to the launch of the spacecraft (with this duration depending upon the complexity and novelty of the spacecraft design). This ensures that operations engineers have enough time to properly assess and address all the potential problems involved in any procedure. Procedures for instrument preparation are typically developed nearer to launch.

All the procedures that are developed make their way into the FOP, which becomes *the* reference for all aspects of spacecraft operations planning. It defines when, where, and how all spacecraft actions will take place. It may be organized into Nominal, contingency, and test procedures.

20.1.1.1 Nominal Operation Procedures

As the name suggests, this refers to all procedures used under normal (or 'nominal') conditions. This includes

- One-off procedures such as may occur during the launch and early orbit phase, e.g. as solar array deployment or perigee boost burns.
- Routine operations that are used on a regular basis throughout the mission, e.g. eclipse monitoring or battery reconditioning.

20.1.1.2 Contingency Operation Procedures

These procedures are used to recover from an off-nominal situation, and in some cases can be highly time critical. An example might include the case where the operator must recover the spacecraft from a situation where all computer and avionics units have switched over to the redundant (backup) hardware following a serious attitude de-pointing (i.e. misalignment of the spacecraft body). In this case, the contingency procedures may be called by an overarching leading procedure or flow chart, which guides the operator to diagnose which unit may have failed and to then select the appropriate procedures required to switch all non-failed units back to their nominal sides.

The hours immediately following launcher separation are a particularly critical time for contingency operations. The operations team must be prepared to respond quickly and correctly to a fault condition, or risk partial or total loss of the mission. For this reason, a wide range of contingency procedures will be prepared which cover all reasonable fault scenarios. Many of these procedures will apply only during this initial launch period and, if not required, will never be used. Nevertheless, their preparation is an important part of the risk reduction process.

20.1.1.3 Test Procedures

Test procedures are devised specifically for spacecraft (or ground infrastructure) test purposes including system verification tests (SVT) that are performed prior to launch, and in-orbit testing (IOT) procedures.

The purpose of such procedures is not to achieve a specific mission objective, but rather to assess the status of the spacecraft (and/or ground systems) and to ensure that the full mission system (spacecraft and ground segment) is operating as expected.

20.1.1.4 Classification by Subsystem

Procedures can also be further subdivided among the various subsystems of the spacecraft discussed elsewhere in this Handbook, including

- Attitude and orbital control subsystem (AOCS), also known as the attitude determination and control subsystem (ADCS)
- Electrical power subsystem (EPS)
- Telemetry, tracking, and commanding (TT&C), also known as telemetry, commanding, and ranging (TCR)
- Payload
- Thermal control subsystem (TCS)
- On-board data handling subsystem (OBDH), also known as the data handling subsystem (DHS). Procedures are very specific to the platform/bus and to the payload of the spacecraft. There is additional variation when operating a satellite constellation, which may comprise similar spacecraft of various ages and levels of design evolution, all being controlled and monitored by the same operations center. Irrespective of any design differences, each individual spacecraft develops a distinct ‘personality’ over time due to anomalies, failures, and other operational constraints (such as fuel depletion, or varying thruster efficiencies). These differences will lead to the creation of further procedures specific to a given spacecraft.

20.1.1.5 Automation

For modern ground control systems, automation of operations may be possible (though not always appropriate). The decision to automate is often about the return on investment. Automation represents a significant investment of effort, which may not be worthwhile for a short-lived mission. However, there may be major economic benefits when considering fleet or long-term operations. Throughout the life of a spacecraft, it is common for the engineers of the flight control team to automate simple, routine operations.

Automation requires that the manually derived algorithms from the (already scrutinized) procedures be translated into computer code. The decision to automate procedures must be taken with careful consideration. There is considerably less margin for error when creating an automated procedure to be executed by the control system

(or the spacecraft). Without human operators, and their engineering judgment, simple mistakes may go undetected. This moves more of the operational emphasis and responsibility to the engineers and analysts (who prepare the automation scripts) and away from the operators themselves. Nevertheless, there are clear benefits to be gained in automating procedures. The human operator is both a filter of, and a source of mistakes. In particular, frequently repeated yet simple operations may be more reliably performed by automation than by a human operator.

Not all procedures will be, or even can be, automated (at least, not without excessive use of resources). Routine operations can often be carried out under automation, and in many cases for newer spacecraft designs, are carried out on-board by the spacecraft control unit (SCU), for example, the switching of heaters in response to temperature variations. The automation can take the form of a small software program uploaded to the satellite, referred to as an on-board control procedure (OBCP). Contingency actions can also be coded for response by both the ground computer systems and the SCU, but it is impracticable for all possible contingencies to be foreseen, and a human/engineering link must remain. Automated procedures may still be monitored by a spacecraft controller as a precaution.

Automation implemented through the on-board software (OBSW) may provide valuable resilience to hardware fault conditions, and reduce the effort needed to operate under nominal conditions. Nevertheless, the added complexity of the control software introduces its own risks, and may drive a need for greater expertise in the operations team to handle problems occurring in the OBSW itself.

20.1.1.6 Procedure Validation

A procedure can be seen as an algorithm. It gives explicit direction based on simple, clear logic. Since all commands that will be sent to the spacecraft should come only from a procedure, the level of detail must be high and strictly adhered to. Any step involving human judgment should be scrutinized to ensure that the exercise of such judgment is necessary. If not, then further development of the procedure should be considered. For this reason, rigorous cross-checking and auditing processes are recommended in order to ensure the procedures are as accurate, complete, and clear as possible.

Procedural validation is a process that can take several years. The genesis of the procedure comes from information derived from the manufacturer’s Spacecraft User Manuals (SUM) or Orbital Operations Handbooks (OOH), other technical documentation and design engineers’ expertise.

Following initial generation of the procedure, the checking and reviewing process must be comprehensive. Procedures must be rigorously tested against the spacecraft simulator (see [Sect. 20.1.5](#)). In addition, as many of the

procedures as possible are also tested on the spacecraft during SVTs or on an engineering model if one is available.

A completed, checked, validated, and authorized procedure is not a ‘final’ product. Procedures must be continually reevaluated for their currency, throughout the lifetime of the mission. Spacecraft configuration changes according to mission requirements, anomalies, failures, ground constraints/problems, as well as countless other planned and unplanned events. Procedures must also evolve in response to changes in the spacecraft’s telemetry and telecommand database (and vice versa).

Validation of procedures may be performed in layers involving evolving, separate organizations. For example, their maintenance may require the participation of the spacecraft manufacturer, the spacecraft operations team, payload instrument experts, or flight dynamics support teams.

This ensures that the procedures are scrutinized from the different points of view that come from the various fields of expertise. Typically, the flight control team will retain overall responsibility for the FOP.

20.1.2 Mission Control Systems

This section describes general features commonly found in Mission Control Systems (MCS). The description that follows is based primarily on the European Space Agency’s Satellite Control and Operation System (SCOS), though most features will have analogs in other commonly used systems.¹

20.1.2.1 General Features

The primary purpose of the MCS is to give an operator the necessary tools to understand the state of the spacecraft, and assist in safely commanding the spacecraft. It is the system through which the operator interfaces with the spacecraft. It displays the telemetry (TM) received from the spacecraft as well as the status of ground station visibility and links. It also provides an interface through which commands are sent to the spacecraft.

The MCS performs a number of functions with the received telemetry

- *Archiving and retrieval*—All TM received is stored, and may be retrieved at a later time.
- *Display*—The status of the spacecraft, as reported through telemetry, is presented to the user. This may be done in simple textual form (name/value pairs), as a

graph, or as a display in which the state of the spacecraft is depicted in a dynamic diagram (e.g. showing switches as opened or closed). Such diagrams are referred to as ‘mimics’.

- *Checking*—Automatic checks are performed on the telemetry parameters in order to assist the operator in detecting anomalous states. This applies to both passive monitoring of the spacecraft and to the confirmation of correct processing of commands.
- *Time Correlation*—Determination of the offset between the spacecraft’s on-board clock and that used on the ground is important for time-sensitive commands (e.g. for the timing of an orbit change maneuver or payload action). The MCS uses timing information contained in the telemetry stream to measure this offset and feed it back into the construction of commands, and the interpretation of telemetry. The MCS assists the operator in constructing and sending commands, also known as telecommands (TC). Important features include
 - *Definition and syntax*—The MCS enforces the construction of well-formed command packets, selected from only those that are defined for the mission, and with all necessary parameters provided.
 - *Time tagging*—Commands may be uplinked for delayed execution, at a predefined time (if supported by the spacecraft). Alternatively, the MCS may support the delayed transmission of a command for immediate execution upon reception. In either case, the MCS should permit the operator to monitor the progress of commands, and to determine what commands are queued for execution.
 - *Pre/post-transmission validation*—The MCS may perform checks against the command, both before and after transmission. Pre-transmission checks can be used to prevent inappropriate (but syntactically valid) commands from being sent. Post-transmission checks can be used to confirm correct receipt and execution of the command.
 - *Grouping of commands (‘stacks’)*—The MCS may support the preparation (plus save and restore) of groups of commands, for rapid sequential (or simultaneous) release. Such stacks may be referred to from procedures. This ensures that complex command sequences can be constructed and checked on the ground, before any of the individual commands are transmitted to the spacecraft. It also permits commands to be executed in rapid sequence, without using time-tagging functionality.

20.1.2.2 Spacecraft Telemetry/Telecommand Databases

The spacecraft telemetry/telecommand (TM/TC) database contains detailed information on every single telemetry parameter and telecommand.

¹ Many MCS implementations are now based upon the internationally recognized standards of the Consultative Committee for Space Data Systems (CCSDS). Adherence to these standards allows interoperability and sharing of infrastructure between operators around the world.

The database serves several purposes. Firstly, it provides the operations team with a definition of the structure of all telecommands (and command parameters) and the location of all telemetry parameters within the downlinked telemetry format/data packets. It will therefore be delivered to the flight control team by the spacecraft manufacturer, and will be used in the preparation of operational procedures (it may be loaded directly by a procedure generation tool in order to aid in authoring). This ensures that all telecommands and telemetry parameters referred to within the procedure are ‘legal’ and well defined.

Secondly, the database will be loaded by the spacecraft control system software in order to enable it to decode the downlinked telemetry and to build telecommands to be uplinked.

The database is also used in the process of writing on-board software, developing mission simulators and defining hardware interfaces.

We now describe some of the common types of information contained within a typical database.

The following information would typically be included, as part of the definition of a telemetry/telecommand parameter

- *Parameter mnemonic*—This is a shorthand code for the parameter, e.g. ‘A292’ (not encoded in the transmitted packet).
- *Parameter name*—An abbreviated name of the form ‘BATT1_CELL21_PRESS’ (not encoded in the transmitted packet).
- *Parameter description*—A brief description of the parameter, e.g. ‘Battery 1, Cell 21 Pressure’ (not encoded in the transmitted packet).
- *Calibration curve*—All parameters require a definition of their encoding method, to allow interpretation of the number that is placed in the field. For example, a binary parameter may use the value ‘1’ to indicate ‘True’ or ‘False’, ‘On’ or ‘Off’, ‘Left’ or ‘Right’, etc. An engineering value range (e.g. 0–0.5 V) may be converted to an integer range 0–100 for encoding.
- *Units*—If applicable, the units of the engineering values will be specified, e.g. volts, amps, radians.
- *Validity conditions (for TM parameters)*—Certain parameters are only valid under certain conditions. For example, the output from a pressure transducer may only be valid if the remote terminal which acquires and processes the pressure transducer output voltage signal is switched on. If the remote terminal unit was switched off then the pressure might read zero, which could be incorrect and misleading. Validity conditions allow such cases to be identified and handled by the control system (e.g. to provide visual indication of a non-valid parameter value).
- *Out-of-limits (OOL) conditions (for TM parameters)*—Most telemetry parameters will be expected to have

values within a certain range to indicate healthy operation of the units to which they pertain. Therefore it is useful to specify safe limits on each parameter so that the operator can be alerted if these limits are transgressed. Two types of limits may be defined—soft (or warning) limits and hard (or alarm) limits. Under normal circumstances, the parameter in question would be expected to have a value somewhere between the soft upper and soft lower limits (‘open’ limit ranges may also be specified—e.g. ‘greater than n’).

A concept of ‘derived parameters’ may also be introduced. These parameters are not downlinked directly in telemetry, but are calculated in the control system from the current values of one or more other telemetry parameters and/or conditions. For example, the state of charge of a battery might not be measured directly but calculated using the output of one or more battery cell temperatures/pressures/voltages/charge currents etc. In this case, a derived parameter can be calculated by the ground system accordingly and added to the information displayed to the user.

Beyond derivation of parameters, telemetry may also be used for trend analysis and fault detection. Appropriate mathematical treatment (e.g. statistical analysis) of telemetry gathered over a long period can provide evidence of equipment degradation prior to an actual failure. Such analysis would typically be performed off-line.

Data exchanged between the spacecraft and ground systems is transferred in blocks referred to as TM/TC packets. Packets must conform to a well-specified structure that allows repeatable encoding and decoding without ambiguity. Telemetry packets typically contain many individual parameters, often grouped into related types or associated with a specific subsystem. For example, a telemetry packet may contain a number of parameters, each representing the temperature in a different part of the spacecraft.

A given packet will reference one or (usually) more parameter definitions, and will also include

- *Packet mnemonic*—This is a shorthand code for the packet, e.g. ‘P123’ (not encoded in the transmitted packet).
- *Packet name*—A short, abbreviated name, e.g. ‘POWER_HK’ (not encoded in the transmitted packet).
- *Packet description*—A brief description of the packet, e.g. ‘Power subsystem housekeeping TM’ (not encoded in the transmitted packet).
- *Type (and subtype)*—A value, encoded within the packet, that allows the structure and purpose of the packet to be determined by the receiving equipment. For telecommands, it may also be useful to define automatic checks that should be performed before and after the transmission of the command.
- *Pre-transmission validation (PTV)*—Also referred to as pre-control, this is a condition which needs to be satisfied

before the command can be dispatched by the ground control system, e.g. only permit a 'mode change' command if the satellite subsystem concerned is reporting an appropriate operational state.

- *Execution verification*—Once a command has been sent it will be necessary to verify its execution by checking telemetry. This may be done by associating the command with a telemetry parameter for which a particular value is expected following execution, e.g. telemetry should report 'ON' following a 'switch-on' command.

A parameter may be encoded in a packet in many ways.

To ensure consistency between the encoding side (e.g. the spacecraft) and the decoding side (e.g. the control system), the method of encoding must be unambiguously defined for all parameters; this is termed value encoding and calibration curves. Such information includes the following.

- *Binary states*—e.g. 'on' or 'off'; 'clockwise' or 'anti-clockwise'; 'present' or 'not present', etc. Such values are normally encoded on a single bit. The calibration curve defines which state maps to '0' and which to '1'.
- *Integer and floating point numbers*—e.g. 'thruster pulse count', 'clock ticks', 'wheel speed', etc. The use of signed versus unsigned encoding must be indicated, as well as the 'width' (or number of bits) used to contain the value. A conversion factor (or calibration) may apply, e.g. 'wheel speed' may operate in the 'engineering' range of 0–4,000 RPM, but be encoded as an integer ('raw') value between 0 and 255 (with an implied loss of accuracy). This permits the relevant information to be encoded with fewer bits than would otherwise be necessary. Note that many methods for encoding integer and floating-point numbers as binary exist, and several may be used within a single spacecraft. The precise method for representing a number in binary form is an essential part of the specification of a parameter.
- *Enumerations*—integers used to indicate one amongst a defined list of non-numerical states, e.g. '1 = Warm-up Mode'; '2 = Standby Mode'; '3 = Off'; '4 = Undefined'. The associated 'calibration' is therefore the definition of meaning associated with each discrete numerical value.

As described, it may be necessary to stipulate conditions for telemetry validity, telecommand pre-transmission validity etc., based upon the current value of various parameter(s) in telemetry.

Validity checking is performed using 'mode equations', which are a combination of telemetry parameter values and logical operators. For example, a detected eclipse condition may imply that normal thermal telemetry limits are inappropriate, i.e.

- IF (battery is discharging) AND (Sun is not present in sensor field of view) THEN (assume eclipse condition and relax thermal alarm limits)

20.1.2.3 Time Correlation

Knowledge of the relationship between the spacecraft's own clock and clocks used on the ground is important for the correct construction and execution of commands, and for the interpretation of telemetry.

It is not uncommon for spacecraft clocks to measure time only in simple terms of 'ticks' (or hardware cycles) elapsed since a defined 'instant'. Say, 10,000 s elapsed since launcher release.

To make use of this 'time' value (e.g. 10,000) the MCS must know

- *Offset*—The 'ground' time at which the 'zero' epoch occurred, e.g. launcher release occurred at 10 h 23 m 30.3 s International Atomic Time (TAI) on 1st January 2012.
- *Rate*—The actual elapsed time per 'tick'. This must be measured by comparing the progress of time, according to the spacecraft, with the progress of time as measured by the ground systems. Inaccuracy in the calibration of the spacecraft clock may result in gradual divergence from 'correct' time.
- *Rate variation*—The rate at which the ground and spacecraft clocks diverge may, itself, vary with time. (i.e. the spacecraft clock tick quickens or slows). The MCS must use timing information in the telemetry stream to determine these values. Accuracy at the level of milliseconds may be required. Nevertheless, many factors affect the time at which a telemetry packet is received by the MCS, all of which must be taken into account when determining the true clock calibration. Such factors include
 - Radio propagation delay from the spacecraft to the ground station (can be many hours for deep space missions).
 - On-board delay between construction of a time stamp (in a telemetry packet) and the actual RF transmission of the packet.
 - Delay in handling and onward transmission by the ground station.
 - Delay in processing at the control center. In order to determine drift rate, and rate variation, several measurements over a period are necessary and synchronization assessment is typically performed as an ongoing activity throughout the mission lifetime.

Note that changing knowledge of (on-board) clock performance, and changing performance of the clock itself, can lead to difficulties in determining the actual time at which an event took place. With a given set of clock characteristics (offset, rate, rate variation), a time stamp will be converted to a certain 'ground' time. However, if future measurements lead to a revision in the values of the offset/rate/rate variation parameters, then this implicitly changes the time at which an early event is understood to have taken place. Thus, it is important to retain information about the clock conversion parameters that are in use at any given time.

20.1.2.4 Time (or Orbital Position) Tagging

The addition of a time-tag (or an orbital position tag) to a command, prior to uplink, can be used to facilitate a range of operator actions that would otherwise be difficult or impossible.

- *Out-of-contact commanding*—Commands may be uplinked during a ground station pass, for execution at a time when a command link will not be available.
- *Critical timing of commands*—Pre-loading of commands to the spacecraft removes risks associated with command link loss (or delay) at the time of execution (e.g. failure or errors in the operation of the control system, network, ground stations, link status, etc.). Once commands are safely loaded, only an on-board failure or deletion command can prevent their execution.
- *Fail-safe commands*—During high-risk operations, commands may be stored on the spacecraft ready to affect a recovery process in the event that commanding capability is (unintentionally) lost. Such commands can be deleted from the spacecraft before their execution time, if operations proceed according to plan.
- *Critical delivery of commands*—For commands which must be executed in a particular order, without omissions, pre-loading provides an opportunity to confirm full receipt, and correct any reception errors (e.g. dropped commands) before time of execution. It also ensures the correct order of execution.

20.1.2.5 File-Based Operations

The preceding description of spacecraft commanding and monitoring is based on the use of application-specific data packets, exchanged in ‘flows’ (akin to Internet packet protocols). Such packets can be regarded as a specialized case of the more traditional computer file.

If appropriately designed, a spacecraft may be controlled entirely by the exchange of conventional computer files. The transfer of such files may then be handled using standard file handling protocols. The spacecraft may be treated as (very) remote network attached storage, just as a networked drive in a typical office IT system. The MCS and the spacecraft communicate by the exchange of files over the network.

Among the advantages of such an approach is the improved ability to take advantage of ‘standard’ commercial IT solutions and technologies, e.g. encryption software, data compression tools, common encoding languages (such as XML), etc.

The use of file-based spacecraft control is well-suited to missions where contact with the spacecraft is intermittent and ‘live’ monitoring is either not necessary or not practical. It is therefore a suitable choice for both low Earth orbit missions (where contact is intermittent and brief) and

for interplanetary missions, where the operator needs to ensure that command stack transfer is complete, error free, and correctly sequenced before execution. For cases where frequent telemetry update is required over long periods, the overheads of file-based transfer may be less efficient than a dedicated packet-based approach.

20.1.3 Mission Planning Systems

Operations planning tools facilitate error free, efficient operation of both the spacecraft and its payload.

Planning tools must be configured (or developed) specifically for each mission, taking account of each mission’s specific constraints and the characteristics of the ground segment, spacecraft, orbit, and payload.

Payload planning tools assist in the scheduling and definition of payload activities, e.g.

- By determining the required spacecraft attitude (orientation) for imaging of an area on the ground, from a given orbital position.
- By optimizing a number of imaging requests to achieve the most efficient execution order.
- By generating (or selecting) command sequences that will be used to execute a payload-related operation, including determination of parameters to load within those commands (e.g. pointing angles, image exposure times, etc.).

Other desirable characteristics in a payload or platform operations planning tool include

- *Encapsulation of operational concept rules and service level agreements*—This includes, for example, ensuring that all time-tagged commands are uplinked a number of contacts (or orbits) ahead of their execution, which constrains contact planning to only those periods when a ground station service is known to be offered.
- *Conflict detection (and resolution)*—Identifying conflicts between requested actions and the state of the system. For example, request for commanding when no ground station (or link) is available; request for uplink or downlink during a contact that is too short to transfer all packets; and request for conflicting attitude changes for observations by different instruments.
- *Resource modeling*—Prediction (modeling) of constrained resources, to ensure capacity to support operational requests. For example, effects on power demand and battery charge state caused by switching of equipment.
- *Translation of high level requests*—Generating low level command sequences, conforming to the spacecraft database, from high level operational requests, e.g. ‘Acquire image of target at location [x, y]’.

20.1.4 Flight Dynamics

Flight dynamics is the term commonly used to refer to the discipline and function of the ground segment element dedicated to satellite trajectory and attitude prediction. A mathematical treatment of the flight dynamics domain is not practical within this chapter. Nevertheless, its importance to spacecraft operations as a discipline cannot be overestimated.

Operational activities requiring input from the flight dynamics function include

- *Determination of spacecraft inertial and performance characteristics*—e.g. mass, center of mass, moment of inertia, drag coefficient and actuator outputs and (mis-)alignment vectors (for example, reaction wheel alignment and torque generated).
- Determination of current orbit parameters, based on observations and/or measurements taken by on-board sensors and ground-based systems.
- *Planning of orbit maneuvers*—e.g. required thrust vector (direction and magnitude/duration) for orbit change.
- *Planning of attitude (orientation) maneuvers*—e.g. required thrust vector (direction and magnitude/duration) for rotation, and determination of any secondary effects on the orbit.
- *Prediction of significant geometrically related events*—e.g. acquisition/loss of signal (AOS/LOS) between a ground station and the spacecraft; conjunction of Sun-spacecraft-ground station line (risking heat damage to a tracking antenna); spacecraft entry/exit from eclipse.
- *Guidance of tracking systems*—e.g. ground station tracking of the spacecraft during a pass; tracking of steerable antenna on spacecraft.
- Generation of orbital files, used as input to the mission planning system (orbit event files) and ground stations (STDMS).

Close interaction is necessary between the spacecraft operations team and the flight dynamics function (where separate). The planning of a maneuver is not a purely mathematical exercise, and involves engineering choices. For example, consider a maneuver to rotate the spacecraft. In three-dimensional space, the operator has a choice of

- *Rotation axis*—Some rotation paths may result in optical sensors being blinded by bright sources such as the Sun, in loss of sunlight on the solar arrays, or in the inability of a steerable dish to remain locked on target.
- *Rotation rate*—A faster maneuver requires greater performance from actuators and risks causing damage to weak or flexible appendages (e.g. solar arrays).
- *Timing*—A maneuver may place the spacecraft at risk and so should be performed at a moment when other risk factors are low (e.g. out of eclipse, or while in contact with the ground).. A range of such considerations will apply to any given maneuver.

Flight dynamics is a broad and detailed topic, and the subject of many dedicated books and papers. The following section provides only an overview of the basic techniques employed for determining the location and velocity of a spacecraft.

20.1.4.1 Orbit Determination

Orbit determination describes the process whereby spacecraft tracking data is obtained and used to model the orbit of the spacecraft. Once this orbit information is obtained, it can be used for the following

- *Orbit propagation*—i.e. prediction of the future location of the spacecraft at a given time.
- *Maneuver planning*—i.e. determination of the required thrust vector to achieve a new orbit.
- *Spacecraft attitude re-pointing/biasing and recovery*—i.e. to provide the spacecraft with a (time-evolving) vector for alignment to the Earth (or other body).
- *Events prediction*—i.e. for prediction of eclipses, sensor blinding, contact periods, etc. In many modern spacecraft [particularly in low Earth orbit (LEO)], orbit determination is greatly simplified by the use of an on-board Global Navigation Satellite Systems (GNSS)—such as GPS.

In cases where GNSS is not (or cannot be) used, orbit determination is achieved by statistical analysis of tracking data from ground stations. This applies to high Earth orbits, and to spacecraft traveling far from Earth. Tracking data may include ranging measurements (distance to target) range rate (velocity along line of sight), and bearing measurements (direction to target).

The distance to the spacecraft is measured by accurate timing of radio signals traveling between the spacecraft and the ground station. Two-way ranging is the most common technique. A continuous, modulated signal is transmitted to the spacecraft. The spacecraft sends a return signal, with a fixed phase relationship (and frequency offset) to the incoming signal. The ground systems measure the time taken for a signal to travel to, and back from, the spacecraft. This time can be multiplied by the speed of light to provide the distance to the spacecraft. Highly accurate measurements of range can be achieved, even over vast distances. Accuracy is tightly coupled to the precision of the timing equipment used to measure the signal flight time. Atomic clocks permit timing measurement in the nanosecond range, equivalent to meter or submeter accuracy.

Instantaneous velocity can be determined from the measurement of signal Doppler shift. The Doppler effect creates a change in received radio signal frequency as a result of velocity along the line of signal propagation. The precision of range-rate measurements is principally governed by the precision with which small changes in signal frequency can be measured, and is not greatly affected by distance to the

target. Here, as with range measurement, use of highly accurate and stable atomic frequency sources is key.

The preceding two measurements, when made from a single ground station, give no information about the direction from the station to the spacecraft (only distance, and rate of change of distance). Direction can be established in one of three ways

- *Directional reception*—A large radio antenna (e.g. several meters) is most sensitive to incoming signals over a narrow angular range (small fraction of a degree). The primary angular range of sensitivity, or beam width, is a function of the antenna diameter and the frequency of the radio signal. Direction to the target can be deduced based upon signal strength within the antenna’s beam. For optical techniques, the principle is the same but can be thought of more intuitively in terms of the imaging resolution and angular magnification of a telescope. While angular resolution does not vary over distance, the absolute positional accuracy that it translates to does. The further a target is from the detector, the less accurately its position is known. For a spacecraft at a distance of 1 Astronomical Unit (au), the approximate distance from the Earth to the Sun, 1 arc-second of angular uncertainty translates into a positional uncertainty of over 2 million kilometers. For this reason, direction finding based on simple beam width is only practical for spacecraft close to Earth.
- *Triangulation*—Combining multiple range measurements taken from different locations (or at different times) can allow more accurate direction finding. A technique known as trilateration ranging combines measurements of range from three stations, each at an accurately known location. Triangulation, based on the targets measured distance from each station, permits its position to be established with considerable accuracy in three dimensions. For the best results, the angular separation of the stations, viewed from the target, must be large. Accuracy is therefore a function of angular separation, precision of range measurement, and positional knowledge of the stations themselves.
- *(Very) Long Baseline Interferometry*—As for triangulation, interferometry relies upon the use of two (or more) stations with a significant angular separation, as viewed from the target. The same signal from the spacecraft is received by both stations. The difference in the arrival times at the two stations is accurately measured and is directly related to the angle between the target and the baseline formed by the two stations. This technique is a key part of the process known as delta differential one-way ranging (D-DOR). As in the case of range and range rate measurements, its precision is tightly coupled to timing measurement and it benefits from the use of atomic timing standards. D-DOR is a key technique for spacecraft tracking in deep space missions.

20.1.5 Simulators

Development of the mission simulator is a crucial part of preparation for any space mission, particularly scientific missions for which orbits, operational activities, and equipment may be unlike those designed for earlier spacecraft.

20.1.5.1 Aim of the Mission Operations Simulator

The mission operations simulator is designed to replicate the actions and reactions of the real spacecraft, when in space. It will be used by the spacecraft operations team, both to develop procedures and to practice them.² The mission operations simulator may greatly simplify a range of physical characteristics (e.g. thermal behavior), where these are not critical to its use in training, procedure validation and ground system testing. Other physical characteristics such as rotational dynamics may need to be accurately modeled.

The mission operations simulator connects to the mission control system, and allows the operator to inject spacecraft commands and receive telemetry reports as if communicating with the real spacecraft, in real time. It fulfills a similar role to that of a flight simulator for airline pilots. In many cases, mission simulators are designed to enable the real spacecraft’s flight software to operate within the simulator, running on an emulated SCU processor. The simulator creates a ‘virtual’ environment, in which the real spacecraft’s control software can execute, as if it was driving the real spacecraft. In fact, it is operating within a simulated SCU, within a simulated spacecraft. Thus, the simulator must be capable of fooling not just the operators but also the spacecraft’s flight software.

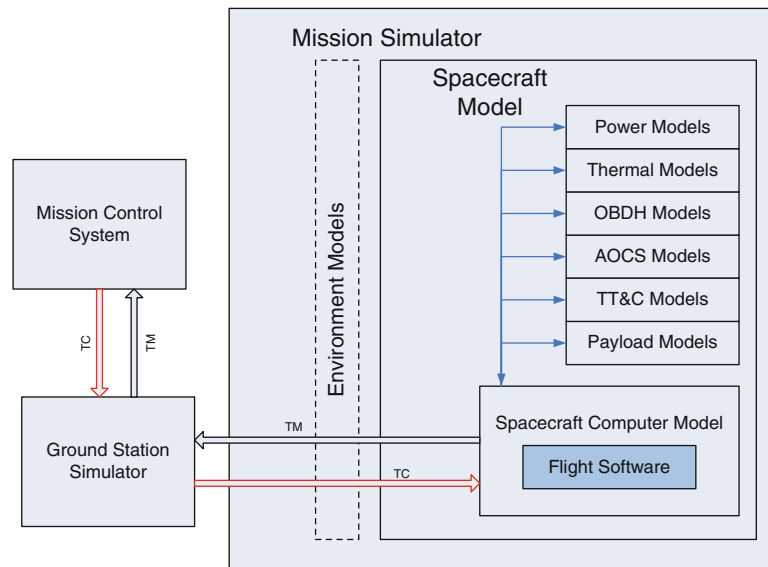
20.1.5.2 What Does the Mission Simulator Do?

An overview of the architecture for a mission simulator is given in Fig. 20.1.

MCS is connected to a simulated ground station. This model receives telecommand packets from the MCS and passes them on to the spacecraft simulator, often via a time delay loop, which simulates the radio signal’s propagation delay (which is operationally significant in deep space missions). Telemetry returning from the spacecraft simulator will be handled in a similar way. It is passed back to the MCS in the form of data received from a real ground station. There is relatively little physical modeling involved within the ground station model—it simply provides the

² A range of other simulators will also be used in the mission development as a whole, focusing on other aspects. For example, the spacecraft manufacturer will employ simulation for structural engineering and dedicated dynamic simulations may be used to validate the on-board software’s attitude control algorithms.

Fig. 20.1 Mission simulator architecture overview



conduit for data traveling to and from the spacecraft. The spacecraft simulator contains models for both internal subsystems and the physical world in which the spacecraft exists. Orbit propagation and communication signal characteristics are both affected by the environment model. The internal modeling of the spacecraft can be seen as being divided into two major areas—modeling of the SCU, with its associated flight software, and modeling of the spacecraft’s actuators, sensors and related hardware. At the heart of the SCU model is the actual flight software of the real spacecraft. In large part, the remaining models exist simply to satisfy the flight software that it is executing within the real spacecraft. The operators, who can only monitor the spacecraft through telemetry packets, base their interpretation of the spacecraft’s state almost entirely on the flight software’s outputs.

20.1.5.3 Modeling Fidelity

The simulator’s training value lies not in flawless physical prediction, but in perceived realism and the encouragement of operator patterns of behavior. In many cases, *accurate* modeling is a requirement—particularly in relation to attitude control equipment and data handling; however, the developer must always keep the goal of the exercise in mind. Too little fidelity and the simulator may encourage bad practice; too much fidelity and cost, and performance and schedules will suffer. The simulator will ordinarily be designed to allow the operations team to validate and practice operations procedures, ranging from first commanding after launcher release, through to the final operational configuration of the spacecraft. In order to achieve this, it is necessary for the simulator to reproduce the behavior of almost every on-board system to some degree.

For example, the subsystems that must be ‘accurately’ simulated include

- *Actuator outputs*—forces, torques, relay positions, etc.
- *Sensor outputs*—voltages, temperatures, rotation rates, etc.
- *Spacecraft dynamics*—attitude control.
- *The visible environment*—Sun, Earth, etc.
- *Electrical power subsystem (EPS) hardware configuration*—switch states, battery charge level, power generation rate, etc.
- *On-board data handling (OBDH) equipment*—TM/TC data formats, remote interfaces, etc.
- SCU
- On-board flight software.

Other systems that may be simulated, with lower fidelity include

- Thermal behavior
- Payload operations (some missions may require higher fidelity than others)
- Physical modeling of power generation—e.g. from solar cells. Inclusion of the real spacecraft’s OBSW within a simulated processor may provide the basis for an objective decision on the meaning of ‘accurate’ modeling. All subsystem modeling must be sufficiently detailed that it fools the OBSW into believing that it is operating within the real spacecraft.

In general, this leads to very precise requirements on the *format* of data passed between the simulated SCU and both the ground and the spacecraft’s modeled subsystems, i.e. the packet standards must be rigorously adhered to.

Another domain in which accuracy becomes critical is in the response of modeled attitude control equipment. The OBSW is constantly looking for any sign that equipment

has failed. Poor modeling of actuator performance, rotation rates or sensor outputs will be quickly detected by the OBSW. This may extend to the modeling of characteristic noise on sensor readouts, if the OBSW expects (or requires) that noise to be evident in the units data outputs.

20.1.5.4 Simulation Campaign

Following the routine use of the mission simulator during the operations development period prior to launch, a simulation campaign or launch rehearsal will often be conducted in the immediate run-up to launch of the spacecraft.

This should be performed under as realistic a set of conditions as possible, using all relevant operations staff in their respective roles. A typical simulation campaign will proceed in three phases, managed by a training or simulations officer

- Nominal flight procedures will be run from the point of launcher separation up until the commissioning stage.
- Procedures will be repeated with various failures/anomalies injected into the simulator such that contingency procedures and fault diagnosis/analysis can be practiced. The nature of the failure may be known only to the simulations officer prior to the exercise, in order to create a realistic training environment. Such failure scenarios are valuable for both technical and human reasons.
- Immediately prior to launch, the simulations officer may then revert to a fully nominal scenario, a dress rehearsal, in order to settle nerves and give confidence to the team.

20.1.5.5 Practical Benefits of Mission Simulators

The mission operations simulator adds value to mission preparations in a number of ways, e.g.

- *Assists the operator in gaining familiarity with the mission control system*—By providing feedback to operator actions, the simulator gives the operator a more complete interaction with the mission control system.
- *Assists the operator in gaining familiarity with spacecraft on-board systems*—The nature of spacecraft operations means that the spacecraft ‘lives in the operator’s head’. A firm mental conception of each subsystem is vital, since there is no physical contact with the object that the operator is controlling.
- *Enables validation of operational procedures, prior to launch or other critical mission phases*—Every procedure that is developed for controlling the spacecraft must be tested using the simulator before it is used on the real spacecraft. Testing is vitally important, to ensure that nothing has been overlooked during preparation. Procedural faults may physically damage the spacecraft or lead to other indirect and unforeseen consequences.
- *Supports interface testing of the ground network and control systems*—The simulator provides a source of telemetry packets and a sink for telecommands. It can be

used to test the complete chain of processing equipment in the ground segment, to ensure that each component is capable of communicating with the next.

- *Supports performance testing of the ground network and control systems*—By producing representative volumes and types of data, the simulator can exercise the ground segment equipment, ensuring that each component processes data correctly, and in a timely manner.
- *Enables validation of the spacecraft’s on-board software*—Executing the real OBSW within the simulator can provide useful feedback on the behavior of the OBSW in a wide range of operational situations. During the development of the spacecraft, the simulator can (and often does) highlight unwanted OBSW features that have not yet been detected on the real equipment.
- *Provides cross-checking of design assumptions and behaviors*—Both the OBSW developers and the simulator developers must make assumptions about the design of the spacecraft hardware and its software. By working independently, the two teams can provide a cross-check on each other’s work—any discrepancies between the OBSW’s behavior and the simulator’s models should be checked. The error could be in either one.
- *Supports training rehearsals for all phases of spacecraft operation*—The simulator is used by the operations team to practice each major mission phase, many times over before the real event.
- *Assists in team-building amongst the operations staff*—Practicing operations leads not only to the development of technical knowledge but also to a better-integrated team. Each participant learns how to work with the rest of the group in order to achieve maximum efficiency.
- *Trains the operations team to cope with in-flight failure scenarios, both technically and psychologically*—The simulator allows faulty hardware to be simulated, and operators to practice recovery techniques. The operator must learn not only the correct technical solution to a problem, but also the correct mental approach to dealing with a potentially stressful, time pressured situation.
- *Develops a wider pool of experts, familiar with the mission and equipment*—Development of the simulator spreads knowledge of the mission design and features. Given the extremely specialized nature of spacecraft operations, such knowledge transfer can be important for the ongoing success of mission control.

20.1.5.6 Ground Segment Test, Verification and Validation

Beyond operations preparation and training, simulation tools also play a role in the validation of the ground segment itself.

Simulation tools may be used as sources or sinks of data flow, with great flexibility. In particular, when integrating a

large ‘system of systems’, the use of simulation-based integration and verification (I&V) systems can support or enable

- Stress testing (i.e. testing the system under maximum load conditions—and beyond)
- Testing prior to system completion (i.e. by using simulated components *in lieu* of real ones)
- Fault testing (i.e. by injection controlled error conditions into the system through the simulator)
- Diagnostic and exploratory testing
- Record keeping for system qualification purposes
- Test case design and configuration control
- System state visualization.

20.1.6 Link Security

This section identifies a number of technologies used for protecting the data link. Erroneous or false signals may be received as a result of either ‘natural’ signal noise or intentional interference.

While most (if not all) missions will incorporate some measure of protection against ‘natural’ signal corruption, more advanced techniques for protection against malicious attack are normally applied only to commercial or defense-related missions, and not to scientific missions.

20.1.6.1 Error Correction

When transmitting binary data, signal noise can result in erroneous reception of the data at the receiving end. The simplest case is that of a binary ‘1’ being received as a ‘0’, or vice versa. The likelihood of such an event increases as the signal gets weaker (e.g. as the spacecraft gets farther from Earth).

Forward error correction (FEC) is a common technique employed to minimize this problem. A detailed treatment of FEC is beyond the scope of this chapter, but the principles upon which it operates can be easily understood. At its most basic, FEC places additional content in the transmitted data to allow transmission errors to be detected and, to a limited degree, to be corrected without the need for retransmission.

Consider an approach which simply duplicates all information transmitted (i.e. the data is transmitted twice). If a random error occurs in reception, the two copies of the data will differ. The receiver can therefore detect the error—though it cannot determine which copy is incorrect.

Transmitting the same signal three times would allow both detection and correction to take place because there is a low probability of the same random error appearing in more than one transmission; hence a majority vote (best of 3) can be applied for each data bit.

In practice, FEC is usually performed using more efficient techniques than simple duplication. Nevertheless, some

measure of signal redundancy is always present, reducing the useable bandwidth on the link. An example of more advanced FEC applied to space data links is Reed-Solomon encoding or Turbo Code, as discussed briefly in [Chap. 15](#).

Further checks can then be performed after decoding to ensure that the data is format-valid, consistent (with itself and/or other data sources), and reasonable (e.g. within an expected range).

20.1.6.2 Encryption

Data encryption is an obvious method for protecting both the command and telemetry links. In the space context, there are some specific issues to address when implementing encryption.

Encryption can add significant volume to the data being transmitted.³ For missions where link data rates are very low, a suitable algorithm should be chosen which minimises this overhead—or alternatively encryption may be entirely omitted.

In the event that a traffic encryption key (TEK) is compromised (no longer secret), it is necessary to replace it with a new private key; i.e. re-keying. This can present a problem for spacecraft since the new key may have to be delivered over the command data link—which is now unsafe. The uploading of new TEKs to the spacecraft may be performed using a second level of keys, which are used only for protection of TEKs themselves. These secondary keys may be referred to as key encryption keys (KEK). This approach relies upon the pre-loading of KEKs to the spacecraft in advance of launch. Only in the event of a TEK change is a KEK required, and the KEK is not required to be transmitted over the data link. This model, based on higher levels of keys for the encryption of lower level keys, can be continued in order to facilitate over-the-air transfer of KEKs themselves, using ‘master’ key encryption keys.

20.1.6.3 Authentication

In some circumstances, the content of a data link does not need to be concealed but must still be delivered in a way that allows the receiver to have confidence that the data is genuine. For example, the delivery of public broadcast data used for safety services must be widely accessible, but guarded against introduction of erroneous data from a non-legitimate source (e.g. signal ‘spoofing’).

This can be achieved by appending the data with a digital authentication stamp, which is the output of a keyed cryptographic hash function. The value encoded in this additional data field is tightly coupled to the content of the main message, and to a secret key. The receiver of the data may

³ It is common practice to generate ‘dummy’ data for insertion in encrypted data flows, in order to conceal the presence (or absence) of true data.

then validate the message by checking, mathematically, that the authentication stamp is consistent with the data to which it is attached. Only a genuine message, authenticated by the correct algorithm, using the correct key, will pass the test.

20.1.6.4 Transmission Security

As an alternative, or an addition, to protection of the encoded data, the manner of transmission may also be used to protect the data link. Pseudo-random frequency hopping is an example of such a technique. The frequency upon which the data is encoded is rapidly changed (perhaps many times per second), according to a pseudo-random sequence. The sequence is the result of a key-based process. To an outside observer, not in possession of the key, the sequence will appear genuinely random, preventing reception of the signal as it jumps from one carrier frequency to another. In fact, the sequence is wholly predictable but only to those in possession of the key.

20.2 Payload Data Ground Segment (and User Segment)

The payload data ground segment (PDGS) is primarily responsible for receiving the science data from the satellite (where applicable), applying the appropriate processing algorithms, and delivering it to the users. Note that the concept of a PDGS relates in particular to observational and scientific satellites, where measurement data is generated by the satellite and must be processed before dissemination to a user community.

20.2.1 Common Payload Data Ground Segment Terms

20.2.1.1 Product Levels

In order to be of practical use to the users, the data received from a satellite must undergo processing by the ground segment. There are a number of defined levels that make up a hierarchical data product scheme.

The data received directly from the satellite is termed ‘raw data’. This is not yet considered to be a data product. The data is reformatted and time ordered, to form the basic ‘Level 0’ data. Level 0 (L0) data is not typically made available to users.

Level 1 (L1) data are engineering products where the data has been converted to engineering units, auxiliary data (e.g. instrument configuration or status information) has been separated from measurements, and selected calibrations have been applied to the data. These products are the foundation from which higher level products are derived. For example, Level 1 radar data may show signal strength

reflected from each point in the imaged area, and variation in surface shape. Increasing degrees of processing may be indicated by sublevels (e.g. Level 1A, 1B, etc.).

Level 2 (L2) data products are typically the same resolution as the preceding L1 data but with further information derived, typically relating to geophysical properties. For example, Level 2 radar data may allow ocean wind speed to be determined based upon the detected wave structure. Level 2 is the most common data product level ordered and used by users, who may proceed to generate further, even higher level data products.

For some missions, further levels of processing are defined indicating additional levels of information derivation from the underlying source data (e.g. Level 3, Level 4). The precise meaning of these terms is dependent upon the mission in question.

20.2.1.2 Auxiliary Data Files

Auxiliary data files are required by the processors within the PDGS to provide the additional supplementary data needed to generate the higher level products (Levels 1 and above). ADFs may be produced by other elements of the ground segment (e.g. attitude steering files, predicted orbit files generated by the flight operations system, etc.), by information generated on-board the satellite (e.g. calibration data not part of the main measurement), or by completely separate bodies (e.g. digital elevation maps, biome maps, other geophysical information etc.). Any data needed to process the higher level products that aren’t direct measurements of the instrument are classified as auxiliary data. The PDGS must establish an interface with the necessary ADF providers in order to complete the processing chain(s).

20.2.2 PDGS Architecture

This section describes the typical architecture of a PDGS for an Earth observation system, as shown in Fig. 20.2.

The output data products of the PDGS are provided to the user segment, for onward distribution to the user community.

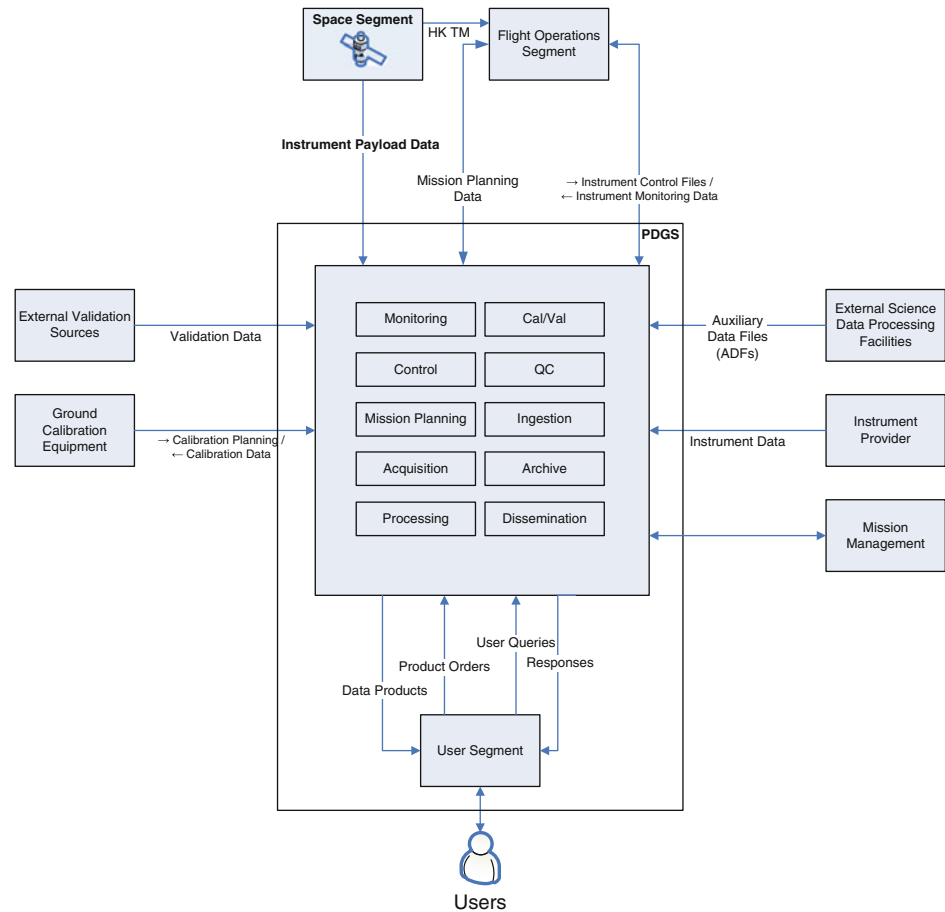
20.2.2.1 Data Acquisition and Ingestion Function

The purpose of the acquisition and ingestion function is to manage the reception of payload data from the satellite and process the received data stream into Level 0 (L0) products.

The acquisition sub-function includes the antenna and RF systems that directly receive the downlinked telemetry from the satellite. Instrument source packets (ISP) are generated for onward transmission to the L0 processor.

The operations carried out by the acquisition and ingestion function are typically located within the ground stations. The processing, archiving and dissemination

Fig. 20.2 Payload data ground segment architecture



functions (see later) do not have to be co-located with the acquisition and ingestion function, and may be geographically dispersed across a number of sites.

20.2.2.2 Processing Function

The L0 processor generates the L0 data from the ISPs produced by the front-end processor. For generation, the L0 products also require certain ADFs made available by the auxiliary data providers.

The L0 data may then be checked for consistency, and any applicable metadata generated and/or extracted. The data, including metadata, is then distributed to the archiving and higher-level processing functions.

The processing function is responsible for processing L0 data up to higher level products. There are two types of processing chain, on-line and off-line; the on-line chain systematically generates the higher level data from the L0 data received from the acquisition and Ingestion function, while the off-line chain is triggered by specific processing requests and receives the L0 data from the archive.

20.2.2.3 Archiving Function

The archiving function receives the L0 products from the acquisition and ingestion function and archives them in a

long-term archive. Archiving strategies for higher levels differ from mission to mission depending upon cost and user requirements, but a common approach is to archive all higher level data (L1 and above) generated by the PDGS (preventing the need to regenerate it from the lower level data at a later time). The function also catalogs all of the data it receives.

The archiving function may be responsible for processing retrieval requests from the user segment, called production orders, for organizing the incoming requests in terms of product type and output media and for scheduling the retrieval of data from the archive for the dissemination function.

20.2.2.4 Dissemination Function

The dissemination and distribution function is responsible for relaying data products to the users following a production order received from the user segment. This production order could take the form of retrieval of archive products, a standing order, or a request for recent acquisitions. This function is responsible for processing a production order, retrieving the ordered products and making them available via electronic means (or where relevant, write them to media) and distributing them to the user.

20.2.2.5 Mission Planning Function

The PDGS planning component is responsible for the generation of the payload plan to be sent to the flight operations segment. This plan focuses on the scheduling of payload activities (e.g. observations or measurements to be performed), and may imply actions for the wider spacecraft (e.g. re-pointing to align on an imaging target). Planning is an iterative process between the FOS and the PDGS, as each has an impact on the other. It is normal for the FOS to provide an initial skeleton plan identifying ‘hard’ planning requirements such as ground station pass availability, required maneuver periods, etc. Payload planning is then included, and a ‘deconfliction’ process then ensues, iteratively, between the FOS and the PDGS planning functions.

The mission planning system includes the following main functionalities

- *Generation and visualization of the payload and downlink plan to be sent to the flight operations segment (FOS)*—These plans are derived from a set of configuration files defining mission objectives: areas of interest, dates, and periodicity of the coverage, instrument modes, downlink stations, recorder properties, etc., that the mission plan must take account of in order to perform the planning activities. These plans may be updated following the receipt of special calibration or observation requests or an identified need for unplanned spacecraft maneuvers.
- *Generation and visualization of the acquisition plan for the stations*—These plans accommodate the planned downlink operations to be sent to the acquisition stations. Note that ground stations used by the FOS and the PDGS for data downlink may not be the same.
- The archiving and configuration control of the mission configuration files and all exchanged data.
- Management of the PDGS and FOS interfaces.
- The mission planning function centralizes the exchange of information between the FOS and the complete PDGS. It may also receive the updates to instrument and processing parameters from the PDGS calibration facility and forward them accordingly to the FOS and to the PDGS processing facilities.

20.2.2.6 Quality Control and Calibration Function

The quality control (QC) and calibration functions are responsible for ensuring that the data products generated by the mission and distributed to the users are fit for purpose. This function covers the systematic and offline quality analysis of data products, the calibration and validation activities and the monitoring of the performance of the instrument.

Systematic QC relates to the activities performed on a routine basis on the data products generated by the PDGS, on either a sample or the complete data set. These can take the form of automated real-time checks within the

processing chain to determine whether a product can be sent to the final user, or the routine expert analysis of the data set to perform mission and product performance analyses.

Offline QC includes the activities of interfacing with the user segment, responding to user queries and identified anomalous products, in addition to the production of quality reports.

The calibration and validation sub-function is responsible for quantitatively defining the system response, both on a product and instrument level, to known controlled signals, and independently assessing the quality of the data products produced by the mission. For Earth observation missions, it is also responsible for geophysical validation activities that independently assess the quality of the geophysical information of the science data (i.e. provision of, and comparison to ‘ground truth’ data, resulting from *in situ* measurements).

Instrument performance monitoring is the sub-function responsible for the monitoring of key instrument performance parameters derived from mission data products, and housekeeping telemetry (HKTm) data from the platform.

20.2.2.7 Monitoring and Control Function

The monitoring and control facilities allow the operators to view the status of each element and facility within the PDGS. Typical capabilities include

- Execution control of processes
- Status monitoring (including warning and alarm states for software and hardware)
- Monitoring of data processing queues, and intervention capability where appropriate.

20.2.3 User Segment

The user segment is responsible for providing a single interface between the mission and users, for a number of services including data ordering and addressing specific problem requests. No other function within the ground segment should have an interface with the user other than the user segment. The segment typically consists of the following sub-functions

- The user order handling sub-function handles user orders and maintains a database of orders and their status. It splits orders according to past (archive retrieval) and future (acquisition tasking), and tracks the status of individual items. It is this sub-function that generates the product orders that are sent to the product order handling sub-function in the archiving function.
- The user ordering application is a client side application that accesses the databases of available data products and allows users to place orders for specific data products.
- The user service catalog provides all archived and potential products as a hierarchy of collections to be ordered.

- The user registration and authorization sub-function handles user account details and provides authorization for certain user services, such as product ordering.
- The user help desk is a first line help desk that directly interfaces with the users on questions and issues with the data or ordering process. It relays any queries that it cannot answer to the QC and Cal/Val function of the PDGS.

20.3 Mission Operations

This section describes significant features of each phase of a mission, from the perspective of the operations engineer.⁴ It begins with brief overview of the effects of orbit on communications links for a range of orbit types.

20.3.1 Impact of Orbit on Communication

The orbit into which a spacecraft will be placed has a major bearing on the way in which the spacecraft will be operated, and the issues that will be faced by the operations team. Three major classes orbit are discussed below.

- *Low Earth orbit (LEO)*—These missions are characterized by brief contact periods with ground stations,⁵ typically only a few minutes long. For polar orbits, placement of ground stations at high latitudes is advantageous because it provides contact up to once per orbit; ground stations at lower latitudes may be able to make contact only every *n*th orbit. High-speed communications links are important to ensure adequate transfer for payload and platform data. Commanding must be pre-prepared out of contact, for (often automated) uplink during a narrow command window. Accurate antenna tracking information must be provided for each pass. In higher orbits (e.g. medium Earth orbit), similar features are present but less extreme. Contact periods are longer, and use of multiple ground stations for (near) permanent contact becomes practical.
- *Geostationary missions*—The fixed line of sight between a ground station and a geostationary spacecraft allows permanent contact to be maintained from a single site. A static (non-steerable) antenna may be used. The timing of operations activities are largely unconstrained by communications considerations, and on-board anomalies can

be detected (and reacted to) quickly. The availability of a continuous data link reduces the need for high data rates on the communications link.

- *Deep space missions*—Communications links may be permanent or intermittent, depending upon the scale of the ground network deployed. With a small number of globally distributed stations (minimum three), continuity of connection is possible, but large dishes are needed for long-range communication. More typically, communication is intermittent but with long active passes lasting several hours. Accurate antenna pointing is required. The time taken for a communications signal to propagate from Earth to the spacecraft (or vice versa) can become large (this is referred to as one-way light time—OWLT). For missions to the outer planets, this period may be measured in hours. Nominal operations and response to on-board anomalies must take account of these long delays, usually accompanied by very low data rates on communications channels. On-board automation for failure detection, isolation and recovery (FDIR) may be required to ensure spacecraft survivability.

20.3.2 Pre-launch Operations

This section, and those that follow, describe the role of the operations team from the earliest phases of the mission through to routine operations (as well as considerations for end-of-life).

The spacecraft operations team are typically involved in reviewing both the design, and the documentation of the ground segment and spacecraft, to ensure that the resulting system is both correctly implemented from a functional stand point, and useable from an operator's perspective. The operations concept will be specified at an early stage and will be a driver for the design of both the spacecraft and the ground segment infrastructure. Early in the mission development, the operations team may consist of just one or two engineers, supported by ground infrastructure experts.

At a later stage, the spacecraft database must be reviewed and debugged, to ensure completeness and correctness. The database will be used as the basis for all spacecraft commanding, and as an input to the mission operational simulator and planning tools. Since development of the simulator in particular must be completed many months before launch, the spacecraft database must be prepared well in advance of this time, as an input to its development.

It should be noted that the control system and the simulator will be developed using the same database. This implies that errors in the database itself may not become apparent, since they will appear in both the simulator and the control system, which will 'agree' on the incorrect

⁴ For further reading, the reader is recommended to study European Cooperation for Space Standardisation (ECSS) paper ECSS-E-ST-70C, describing ground systems and operations for space systems.

⁵ At the time of writing, most LEO missions continue to use direct space-to-ground communications links, rather than relay links via other communications satellites.

definition. Nevertheless, the processes of encoding and decoding should have been independently implemented, thereby giving confidence in the algorithms of the control system to encode and decode. Further independence is achieved in reviewing the correctness of the database, through the development of the simulator. The production of simulator models requires that the developer pays detailed attention to each item in the database, and runs operational simulations in scenarios not yet testable with the real spacecraft. There is an improved likelihood that either the operations team *or* the simulator development team will identify errors in the database.

Once the control system is well developed, the by-now expanded operations team will dedicate much effort to activities such as

- Writing/reviewing/testing operations procedures (see [Sect. 20.1.1](#))
- System verification test (SVT)
- Defining mode equations (see [Sect. 20.1.2](#))
- Defining derived parameters (see [Sect. 20.1.2](#)). SVTs involve connecting the live spacecraft to the ground control center, whilst still ‘on the ground’. This is usually achieved using wide area network (in some cases the spacecraft may even be in a different country to the control center). A complete end-to-end test using specially designed procedures is then performed. The ability to correctly command the spacecraft using the control center is exercised, which may expose problems where the simulator inaccurately modeled the spacecraft’s real behavior. There will be certain limits to exactly what spacecraft functions can be commanded in these conditions (thrusters will not be operated inside a clean room, for example). Additional support equipment may be needed to verify operation for some subsystems (e.g. a display panel to show which thrusters are being commanded, without actual operation of the thrusters).

During the final period before launch, operations rehearsals will be conducted under simulated conditions. Refer to [Sect. 20.1.5](#).

20.3.3 Launch and Early Orbit Phase

The launch and early operations phase (LEOP) refers to the first days of a mission from launch up until the point when the platform has been established to be in good working order. It includes

- Preparation for launch
- First contact with the spacecraft
- Monitoring/performing initialization sequences
- Achieving the final operational orbit/trajectory
- Performing check-out of platform (sub)systems.

20.3.3.1 Preparation for Launch

During this period, a TM/TC link and power supply is maintained with the spacecraft via an umbilical link, as the RF equipment and solar arrays (if applicable) will be stowed inside the launcher fairing. Flight software is loaded and booted up under ground control and it is usual for operations staff to monitor the health of the spacecraft for unusual telemetry, particularly with regard to the thermal control subsystem and the batteries’ health and state of charge. The umbilical link is removed shortly before launch.

20.3.3.2 First Contact with the Spacecraft

This is the section of a mission immediately following the separation sequence (i.e. release from the launch vehicle) and is the point where the operations team will take over control of the spacecraft.

Directly after separation, an initialization sequence will be performed in order to switch on the necessary spacecraft systems for basic operation and initialization of the correct software control modes. The most important operations at this stage are to establish a communications link, and to ensure adequate power supply (e.g. from solar arrays). Commonly, essential actions for these tasks will be performed by an automated on-board software sequence. This ensures that a delay in establishing a command link (or dropout of an established link) does not unduly endanger the spacecraft’s chances in the critical first hours of the mission.

20.3.3.3 Monitoring/Performing Initialization Sequences

Once a space-to-ground communications link has been established, operators can then continue with the initialization sequence. This can be performed manually by telecommand, but more commonly it will be performed autonomously by software, in which case the operations team will simply monitor telemetry (with specialists dedicated to each spacecraft subsystem). If problems are encountered during the sequence, it will be aborted and a backup plan of contingency procedures adopted to perform the initialization sequence manually, by telecommand.

Actions will include the initialization of the reaction control system and activation of the relevant attitude sensors for this stage of the mission. Once this is done, a stable attitude (orientation) can be established. In most cases, this means reaching (and maintaining) an appropriate alignment with respect to the Sun. Spacecraft using deployable solar arrays as the main electrical power source will perform (at least partial) deployment of the solar arrays at this time. One reason that the arrays may only be partially deployed (e.g. outer-most panel sections only, per wing) is that high

thrust/torque maneuvers during final orbit acquisition can exert damaging torques on the extended array structure, and on the solar array drive mechanism at the base of the arrays.

20.3.3.4 Achieving the Final Operational Orbit/Trajectory

If the launcher delivers the spacecraft into a transfer orbit, then it is necessary to acquire the final orbit or trajectory by a series of maneuvers. This can include orbit raising, orbit circularization, plane changes, longitude drifts, or injecting an escape change of velocity, ΔV , for deep space missions.

Small thruster burns are used to calibrate the response of the spacecraft's accelerometers and gyroscopes, before undertaking larger orbit changing maneuvers.

During this period spacecraft visibility from mission ground stations will need to be predicted based on the measured orbit, and TM/TC links will then be maintained by switching between the ground stations.

In the case of geostationary missions, after separation, the spacecraft will usually be in geostationary transfer orbit (GTO) or super-synchronous transfer orbit (SSTO), both requiring a series of thruster burns to circularize the orbit at the final height. Before each engine firing, various sensors will be used to accurately acquire the correct spacecraft attitude, effectively selecting the direction in which the spacecraft will accelerate.

Once a geostationary orbit is acquired, the required longitude position is achieved through east–west maneuvers. See [Sect. 20.3.5](#).

20.3.3.5 Configuring the Spacecraft

During LEOP, full deployment of the solar arrays will be performed, as well as any antennas/reflectors, etc. This will usually involve activating various pyrotechnic devices, including electro-explosive devices (EED) or pyro bolts, wire cutters, (e.g. thermal knives), mechanical release mechanisms, winder motors etc. This is a critical period as failures of pyro units can severely (even fatally) affect a mission.

In-orbit testing (IOT) is carried out, characterizing the responses of all the on-board units and identifying any faults or unexpected behaviors. In order to minimize the amount of unnecessary switching, it is common to initialize first the redundant (back-up) equipment and, once that functionality has been established and characterized, reconfigure to use the prime ('normal') equipment.

20.3.4 Commissioning

For a complex payload, which may consist of hundreds of identical switches and power amps etc., the 'switch-on' process can take many hours or even days.

The approach to commissioning a given payload will strongly depend upon the nature of the payload. In general, the aim is to confirm that all elements of the payload are able to be switched on and that no faults are detectable. Since hardware failures often occur at the moment of power-up, it is advisable to minimize the number of power cycles that any equipment is subject to. For this reason, commissioning tests are sometimes performed on the 'backup' hardware (or 'B' units) first, so that it can then be switched off and left off. This avoids the need to take the primary hardware (or 'A' units) through a power-up-down-up cycle at the beginning of the mission.

For most payload types, some kind of calibration activity will be required. Calibration is a process of measuring the outputs of the unit against known signals, in order to determine any offset or bias that is present in the output. This can then be adjusted for, either by tuning the instrument to give the correct output or by changes to the processing of the data at a later stage.

As an example, the calibration of an Earth observation spectrometer payload may require the collection of spectrum data locally, on the ground, to confirm that the spacecraft sensor's output matches 'ground truth'.

The calibration activity may be repeated throughout the lifetime of the mission, if necessary, to ensure that the instrument remains as accurately calibrated as possible.

20.3.5 Routine Operations

This is the period of the mission that is begun once the spacecraft has been commissioned and is performing the tasks for which the mission was designed. However, missions can vary greatly in purpose and scope and hence the term 'routine operations' can mean different things.

20.3.5.1 Geosynchronous Missions and Station-Keeping

Geosynchronous spacecraft orbit the Earth with a period of 1 sidereal day, and hence their sub-satellite points maintain the same longitude. Geostationary satellites are a subset of these, where the inclination in relation to the equatorial plane is also controlled by keeping it at or close to zero. These two terms are often (and erroneously) used interchangeably.

Geosynchronous spacecraft are relatively straightforward to operate. Since the spacecraft remains at a fixed (or nearly fixed) location in the sky, the task of establishing a TM/TC link is greatly simplified. In this case, it is often possible to maintain near-24-hour communications with the vehicle.

The network of ground stations can be reduced to a single site and an essentially fixed antenna. There is no need to predict ground passes or to limit operations to a finite

commanding window. Nevertheless, geostationary spacecraft do drift off-station and require regular orbit maintenance or ‘station-keeping’ maneuvers.

Due to the nature of the Earth’s non-spherical gravitational field, there are stable and unstable locations for geostationary orbits. At a stable node (approximately 75°E 105°W), orbit correction maneuvers are required only rarely every few months.

In all other locations, the spacecraft will tend to drift in a consistent direction as time goes by. Since geostationary spacecraft are required to remain within a specified ‘box’, the operations team must periodically correct the orbit. Orbital maintenance burns may occur on a weekly basis.

Inclination of the orbit will result in a daily sinusoidal motion of the satellite in a north/south direction about the equator. The degree of north/south movement is directly related to the inclination of the orbit.

The level of inclination that can be tolerated is a mission design parameter. From an Earth-based observer’s perspective, inclination (when combined with eccentricity) will manifest itself as a ‘figure eight’, or lemniscate motion of the spacecraft in the sky over the course of a day; see also [Sect. 4.4.3](#). A large degree of motion may introduce complications for communications links, both for the spacecraft and for communications service users on the ground. The motion may be accommodated by either a wide gain pattern on the antennas or by steering mechanisms. A tightly maintained inclination (at or close to zero) requires regular maneuvers—e.g. on a fortnightly or monthly basis.

Maneuvers are generally split into the following categories

- *East/west station-keeping maneuver*—Trimming of spacecraft’s longitude.
- *North/south station-keeping maneuver*—Control of drifting orbital inclination (only applicable to geostationary satellites).
- *Eccentricity correction*—Periodic circularization of the orbit. This is often combined with station-keeping maneuvers.
- *Attitude maneuvers*—Dumping excess momentum that was acquired through external sources of torque (e.g. solar pressure, gravitational torque, etc.). Different missions apply different philosophies for keeping the spacecraft within their assigned ‘box’. Some spacecraft only perform maneuvers if they are about to move outside of a defined dead-band (on the order of 0.1–1° wide). The maneuver will take the spacecraft all the way to the other side of the box, in expectation of it slowly drifting back through the ideal location. This approach minimizes the number of burns that need to be performed. Other missions try to maintain the most accurate location possible at all times. This can lead to a much higher frequency of maneuvers.

For the majority of the year, the spacecraft is never in shadow owing to the orbit’s inclination (see [Sect. 4.4.3](#)) with respect to the ecliptic plane. During two short periods every year, around the vernal and autumnal equinoxes, geosynchronous spacecraft enter ‘eclipse season’ where they undergo daily eclipses. This gives rise to the following concerns:

- *Orbit/attitude effects*—For most of its life, the spacecraft is continually exposed to solar radiation pressure. During this short period, the spacecraft undergoes short period of darkness, which will impact on predictions of orbit and attitude evolution.
- *Reduction in power generation*—For communications spacecraft, power demands can be very high. The operations team must ensure that the life of on-board batteries is maintained for as long as possible. Battery conditioning operations must be executed to ensure that the batteries retain the ability to store their design capacity.
- *Sensor blinding and Earth tracking loss*—The periodic, transient presence of the Sun within the field of view of attitude sensors can give rise to physical damage and/or ‘false’ Earth detections. See [Sect. 20.3.5](#).

20.3.5.2 Sensor Blinding and False Targets

Spacecraft using optical (including infrared) sensors for the maintenance of attitude are susceptible to guidance errors caused by the presence of unexpected astronomical bodies in the field of view.

As an example, in geostationary spacecraft, the use of infrared-based Earth sensors is common. These sensors detect the thermal energy of the Earth and, by monitoring the strength of the detected signal, permit the spacecraft to maintain an Earth-pointing attitude. The presence of other heat sources in the bore sight of the sensor can lead to false ‘Earth’ detections and errors in attitude control. Both the Moon and the Sun present a risk in this context.

From geostationary orbit, the Sun will appear to move around the spacecraft once per day. Each day, it will pass above, behind or below the Earth from the perspective of the spacecraft. During the periods around the equinoxes, the Sun will pass behind the Earth. It is these periods close to the equinox that give rise to significant operational risks for satellites guided by infrared Earth sensors. As the Sun moves close to the Earth’s limb, the sensor may begin to track off the Earth and follow the Sun, rather than the Earth. This will rapidly lead to a loss of correct spacecraft attitude.

The transitory presence of the Sun (or Moon) in the field of view of the sensor must be accommodated in either the design of the sensor or the operational procedures of the spacecraft. For example, the sensor (or a subfield of it) may be temporarily disabled by the operations team or on-board control software, when the presence of the Sun (or Moon) is either expected or detected.

Similar technical or process-based approaches should be applied to the use of other sensor types such as star trackers, magnetic field sensors, Sun sensors, etc., wherever the presence of a transitory input may give rise to guidance errors. Such events can occur, even in deep space missions (e.g. presence of an asteroid or Sun in a star tracker field of view). The operations and flight dynamics teams must take precautions to avoid such events or minimize associated risks.

20.3.5.3 Orbit Maintenance

Approaches to station-keeping and orbit maintenance in Earth orbit will vary depending on the requirements of the mission.

For geostationary missions, refer to [Sect. 20.3.5](#).

For navigation systems (e.g. GPS, Galileo), precise knowledge (and prediction) of the satellite's orbital path is central to the delivery of the service. Orbital tracking is performed over long periods (weeks) in order to derive the most precise estimates possible of the future path of the spacecraft. This information forms the basis of the delivered navigation service, allowing users to determine their own position relative to the 'known' location of the satellites. As a result, knowledge of the satellite's orbital path is more important than control. Maneuvers might be performed only a handful of times over the complete lifetime of the satellite, since each maneuver introduces new variability to the future orbital path.

For missions in LEO, atmospheric drag becomes significant. The satellite is slowed by the atmosphere, bringing it lower, thereby further increasing the drag effect. Routine orbit raising maneuvers are required to prevent eventual atmospheric reentry. Often, satellites in LEO are also required to maintain a specific ground track, passing over the same points on the ground in a predictable pattern. Further maneuvers may be necessary on a routine basis to ensure that the ground track does not drift with time owing to gravity perturbations similar to those that affect geostationary satellites.

20.3.5.4 Solar Sailing for Attitude Control

Atmospheric drag and solar radiation pressure can give rise to attitude perturbations (unwanted rotation) on the spacecraft. In the short term, these external forces may be compensated for by reaction wheels,⁶ which 'absorb' the additional angular momentum, allowing the spacecraft body to remain rotationally still. Nevertheless, if the external force is applied in the same direction over a long period, reaction wheels will continue to accelerate, eventually

reaching the design limit of the equipment. It is then necessary to 'dump' the accumulated angular momentum from the spacecraft. Often this is achieved using reaction control thrusters or, in LEO, by the use of magnetorquers.⁷ Both methods require some expenditure of resources—fuel, in the case of thrusters or power in the case of magnetorquers.

A passive method is available for attitude control based upon the use of solar pressure and steerable solar arrays. Appropriate angling of the arrays, will result in a directed torque on the spacecraft—in a similar manner to wind on a ship's sails. Although the force generated by solar pressure is small, it can be consistently directed such that the effect accumulates over time. For geostationary spacecraft, this entirely passive momentum dumping technique may significantly reduce the overall mission requirement for propellant, which would otherwise be needed for thruster-based momentum dumping.

20.3.5.5 Propellant Management

Determining the quantity of propellant remaining in the spacecraft's tanks requires indirect methods. Since the spacecraft is in a microgravity environment, the tank cannot be 'weighed' and fuel cannot be assumed to be at a specific end of the tank unless it has been physically forced there.

Methods available for determining the remaining quantity of fuel include

- *Calibration maneuvers*—A thruster burn, of known total impulse, can be used to determine the total mass of the spacecraft. If the dry mass is known, then the quantity of fuel remaining can be derived. This method requires good calibration of the thrusters, since the applied impulse can only be determined from estimates of the firing time and the thruster performance.
- *Thermal Capacity*—Applying heaters to the tank results in an increased temperature. The rate at which the temperature rises is related to the quantity of fuel that is being heated, and to the heater power that is being applied. The (time-varying) dispersal of fuel throughout the tank will impact the rate at which heat reaches different locations within the tank. Heating over several hours may be needed to achieve a dependable result.
- *Dead-reckoning (pulse counting)*—The total fuel consumption can be estimated based upon the known accumulated firing time of the thrusters (for both maneuvers and attitude control). This method assumes a well-calibrated relationship between firing time (or number of thruster pulses) and the volume of fuel consumed. The effect of reducing the tank pressure over the mission

⁶ A motor-driven, spinning mass. Rotational acceleration and deceleration of the mass produces a torque that can be used to steer the spacecraft body or to resist an external torque. Wheels can also provide gyroscopic stability.

⁷ Devices for generating a controlled magnetic field, with a known orientation. Magnetorquers can be used to apply a control torque on the spacecraft, when a significant external field is present.

lifetime must be taken into account when considering the estimated fuel flow rate.

Fuel gauging is an important consideration. Calculation of the thruster firing required to achieve a certain change of velocity, ΔV , depends upon knowledge of the total spacecraft mass, including the remaining fuel. Fuel depletion is also commonly a limiting factor on the operational lifetime of a satellite. If the satellite is to be deorbited (or placed in a ‘graveyard’ orbit) then the terminal maneuver must be conducted while sufficient fuel remains on-board. In the case of a geostationary satellite, this maneuver, known as a graveyard maneuver, involves boosting the orbit height by about 300 km—a requirement recommended by the International Telecommunications Union (ITU). This ensures that the primary geostationary belt remains usable for new missions.

An incomplete terminal burn may be worse than no burn, because it will place the satellite in an uncontrollable and eccentric orbit that may intersect the paths of other operational spacecraft—or produce an unpredictable reentry profile.

An unnecessarily early terminal maneuver reduces the value of the mission, either in scientific or commercial terms.

Note that for spacecraft with multiple tanks, periodic ‘rebalancing’ of fuel may be required to prevent large changes in the overall center of mass of the spacecraft as tanks deplete asymmetrically.

20.3.5.6 Battery Reconditioning

Battery reconditioning is the process by which a battery’s ability to efficiently store charge is maintained by periodically subjecting it to a deep or complete discharge through a resistive load (sometimes a heater is used) and then recharging back to nominal operating charge levels. This can be performed manually or sometimes under software control, but essentially the decision of when to perform this operation and on which battery is an operator decision.

Reconditioning is made necessary due to the hysteresis effect on battery charging that is a result of aging. For example, nickel–cadmium batteries contain two forms of nickel hydroxide: a β -form, which is electrochemically active (and hence discharges first) and a γ -form, which is relatively inactive. Over time, the β -form is converted into the γ -form. A high ratio of β -form to γ -form is needed and this can be achieved by periodically conditioning the battery.

Geosynchronous satellites are subject to two eclipse seasons per year in which eclipses of the Sun by the Earth occur every day (some lasting over an hour) for a period of several weeks. This results in significant battery discharges, forcing the electrical power subsystem (EPS) to restore main charge levels on the batteries once the Sun is regained on the solar arrays. It is a commonly employed strategy to

recondition a battery shortly before an eclipse season, both to improve its efficiency and to characterize any problems or degradations of battery cells etc. Typically, one battery will be reconditioned before the vernal equinox eclipse season and the other will be reconditioned before the autumnal equinox eclipse season. Rarely would both batteries be reconditioned during the same season, and never at the same time. This avoids any risk of simultaneously damaging both batteries (or connected systems).

20.3.5.7 Thermal Management

Much of the control of spacecraft heaters will be performed autonomously either by using heater control loops or mechanically by using thermostats. However, it may also be necessary to manually control heaters for a variety of reasons.

Thermal modeling before launch is not perfectly accurate and will only have been approximated in spacecraft simulators. Adjustments to heater loops and heater statuses during the early life of a spacecraft are normal. Manual switching of replacement heaters is also performed, either when platform or payload units are switched intentionally or in response to anomalies and unit failures.

Manual switching of compensation heaters may become necessary at particular points in the spacecraft’s lifetime. For example, degradation of the exterior surfaces may lead to changes in reflectivity/absorptivity which necessitate heater switching. Similarly, the performance of certain electronic units may degrade over the mission, leading to changes in the amount of heat that they dissipate.

In such situations, if heater loops are under software control then it may be necessary to set new thresholds and filters for particular thermal control loops, if this function is available to the operator. Depending upon the spacecraft, this could be performed either with dedicated thermal control configuration commands or via a patch of the thermal control software (or both).

By means of example, the thermal environment of a geostationary satellite can be subject to the following factors

- *Daily Changes*—The $-X$ (Earth-facing), $+X$ (azimuth), $-Z$ (west) and $+Z$ (east) faces of the spacecraft will receive sunlight at angles of incidence between -90° and $+90^\circ$ on a daily basis over a period of 12 h at 12 hourly intervals. At other times they face deep space.
- *Seasonal Changes*—The $+Y$ (south) and $-Y$ (north) faces receive sunlight at angles of incidence between approximately -25° and $+25^\circ$ over a 6 month period at 6 month intervals such that at an equinox no sunlight reaches the north and south faces, at the summer solstice the north face receives sunlight, and at the winter solstice the south face receives sunlight. The state of high power payloads can have a large effect on the thermal equilibrium of the

spacecraft. For communications satellites, the thermal budget will typically assume that the payload is continuously switched on and dissipating a large amount of heat. During periods when the payload is switched off, ‘simulation’ heaters may be required in order to substitute for the heat normally produced by the payload.

20.3.5.8 Non-geostationary Earth Orbits

From an operations perspective, non-geostationary spacecraft in Earth orbit must deal with a greater number of issues than their geostationary counterparts. In addition to coping with the effects of eclipses (which may occur far more frequently, depending on the particular orbit), the operations team must be able to cope with the variation in satellite visibility from a given ground station.

For non-geostationary missions, prediction of spacecraft visibility from ground stations becomes a more demanding task. If only a limited number of non-ideally situated ground stations are available, as is usually the case, then there may be periods of many hours when the spacecraft is out of contact. The operational impact of such limited visibility is that any commanding must be completed within a finite time, before the uplink is lost again. The ground station must be accurately commanded with a predicted pointing direction to acquire and maintain the spacecraft until visibility is, once again, lost.

Periodic loss of contact has significant implications for the way in which operations are conducted. Any task that creates a risk to the spacecraft (e.g. maneuvers, changes of control mode, switching of equipment) must be very carefully planned, so that a safe state is assured when the link is lost.

It is for this reason that many spacecraft are designed with a timeline or master schedule function on-board. See Sect. 20.1.2.

Note that simple geometric visibility from ground stations to the spacecraft is not sufficient to maintain a viable command link. As the spacecraft approaches the horizon (either skimming along it or descending below it), signal strength begins to drop, noise increases, and objects may begin to block the signal (hills, buildings, etc.). Thus, the exact moment of terminal signal degradation may be subject to some uncertainty. It is normal practice to choose a ‘safe’ elevation angle (e.g. 5°) that can be relied upon to provide a strong link. The predicted duration of the pass will be based upon time of ascent (acquisition of signal, AOS) and descent (loss of signal, LOS) past this angle.

20.3.6 Deep Space Missions

The nature of deep space missions places greater demands on all aspects of the program, from beginning to end.

Operational considerations include

- *Experimental hardware*—Most deep space missions rely on specialized or novel equipment to achieve their goals. With new equipment comes the need for additional training, specialist system knowledge, and risk reduction planning.
- *Maintaining a command link*—The greater ranges involved impose strict limitations on the link budget. Establishing a viable link through an omni-directional low gain antenna (LGA) in orbit around Mars requires a very large receiving dish on Earth.⁸ For a high gain antenna (HGA), accurate pointing is needed. In addition, prediction of spacecraft visibility becomes more complex. For example, Mars Express can only communicate when Mars is visible from the ground station *and* when Mars Express is not hidden behind Mars. In addition, the spacecraft itself may have a steerable antenna (e.g. the European Space Agency’s Rosetta mission) or a fixed antenna (e.g. ESA’s Venus Express mission). Flight dynamics must provide the correct inputs to guide the antenna (or the spacecraft body) such that a link can be established. Contact passes for deep space missions are typically longer than those for LEO missions. The spacecraft remains visible for several hours at a time, driven by the rate at which the Earth rotates, carrying the ground station out of view of the spacecraft. Since data rates are low for deep space links, such long passes are vital to achieve sufficient time for data transfer (uplink and downlink).
- *Orbit determination*—Over the extreme ranges of missions like Cassini-Huygens (to Saturn), measurement and prediction of exact orbital parameters is a non-trivial task. Such data is required to predict periods of link loss, local eclipsing, maneuver timing (and vector), observation timing and radiation environment.
- *Light-time*—At interplanetary ranges, propagation of radio signals takes a significant period of time. Operators may have to wait many minutes, to hours, before seeing the results of a telecommand. As a result, the level of autonomy on the spacecraft is high since mission control is unable to act quickly in the event of a fault occurring. Commanding is done almost exclusively via the on-board timeline, and often using on-board control procedures. The time taken for a signal to travel between the Earth and the spacecraft is referred to as one-way light time (OWLT).
- *Solar conjunctions*—Interplanetary missions may pass through periods of several weeks when it is not possible to communicate with the spacecraft. When the Sun comes between Earth and the spacecraft, all operations must halt until a link can be reestablished. Operators must configure

⁸ Following a serious failure, the spacecraft may reorient itself towards the Sun for maximum power generation. An LGA link may be required, if the spacecraft has lost tracking of the Earth.

the spacecraft in a ‘reduced activity’ state prior to link loss, and then reestablish normal operations on reacquisition.

- *Ground networks*—Deep space missions require particularly large antennas, of which relatively few are available worldwide. Often, the same dishes will be shared by several missions—reducing the available uplink time for each. Considerable effort goes into optimizing the use of facilities like NASA’s Deep Space Network and ESA’s ESTRACK; see Sect. 2.2.1.
- *Non-routine operations*—For most commercial satellites, routine operations begin after a few weeks or months of in-orbit checkout. In the case of interplanetary missions, the mission may continue to evolve over very long periods, requiring sustained planning activity and application of specialist knowledge. Missions such as Cassini-Huygens (to Saturn) or Rosetta (to a comet) enter something approaching ‘routine’ operations only after many years of flight. The goals and tasks of the operations team change constantly, as different phases of the mission pass and in response to new scientific objectives.
- *Changing environments*—Different phases of the mission may be subject to widely varying environmental conditions. ESA’s Rosetta comet lander must cope with variations in the solar environment from Earth’s orbit to the dimly lit, cold beyond the orbit of Mars and the dusty space surrounding an active comet. BepiColombo must move from Earth’s orbit to the intense, hot environment of Mercury. These issues are not just a problem for the spacecraft designer. Modeling and prediction goes on throughout the lifetime of the spacecraft, to allow mission operators to make the best decisions in unforgiving situations (for example, to determine the level of power that will be available from solar arrays, and spacecraft temperature, as distance to the Sun varies).
- *Changing technology*—Most deep space missions are many years in the making. Allowing for changes in technology is an important part of mission operations planning. Equipment and software on the ground must be maintainable for the whole lifetime of the project. ESA’s Rosetta mission can trace its origins back to 1993, with launch in 2004, arrival at its target (a comet) in 2014, and completion of its mission in 2015. As an example, during this period the widely used Microsoft Windows operating system has evolved from Windows 3.1 through to Windows 8 (at time of writing in late-2013), and will no doubt evolve further by 2015.

20.3.7 On-Board Software: The Operator’s View

The software that runs on the spacecraft’s on-board computer provides the main interface through which the

operator drives the spacecraft. The vast majority of telecommands and telemetry packets are issued to, or received from, the spacecraft’s on-board software (OBSW).

Data packet exchange with the OBSW is the primary mechanism by which the operator is able to manage

- On-board data stores
- Attitude and orbit control, and associated sensors and actuators
- Payload activities
- Power storage and use
- Thermal conditions. For the operator, a strong understanding of the OBSW at the architectural level is vital. In earlier generations of spacecraft, commands would be specific to particular actuators or sensors, e.g. ‘Close power relay for reaction wheel 1’. While these commands are normally still available to the operator in modern spacecraft, the design of OBSW tends to provide layers of abstraction, or to allow for more ‘goal’ based commanding. For example, the operator may issue a single ‘Enter Sun pointing mode’ command. This command will be implemented by the OBSW through a complex series of actions, control loops and decision processes. The net result may involve many sensors, many actuators, much computation and a considerable period of time to complete—but from the operator’s perspective is a ‘simple’ change of mode.

In general, the operator is able to monitor the progress of actions on the spacecraft solely through the information fed back in the form of telemetry packets.⁹ Hence, involvement in the design of the OBSW is an important aspect of the operator’s role in the mission. As with any complex machine, a spacecraft must be designed with operability in mind. The parameters available in telemetry are, to a large extent, fixed before launch. The operator must be confident that they have the ability to fully understand the state of the spacecraft, in both nominal and failure conditions, based purely on interpretation of those telemetry packets and/or fields that have been ‘built in’. The implications of this dependence upon data parameters can be surprising. For example, the unfolding of a solar array is an action that may not be directly observable by the operator. Its success can only be indirectly deduced by its effects on other systems—e.g. increasing voltage as the Sun illuminates the arrays, a change in the spacecraft’s moment of inertia (itself indirectly observable through subsequent motion of the spacecraft), and the physical closing of a contact switch as the array moves. The spacecraft operators and designers must foresee the need to monitor a characteristic of the spacecraft, and then ensure that a method is available to do so.

⁹ In the case of maneuvers, externally observed flight dynamics data (e.g. observed acceleration) may also provide an important route to monitor progress.

Similarly, for commanding, it is important to ensure that a command is available in the OBSW for every task that the operator may want to undertake, in both nominal and failure scenarios. Here, again, the operator has a significant role to play.

The end effect of issuing a command to the spacecraft is dictated by the design of the software routines that manage that command on-board. An operator should not issue a command to 'Enter Sun pointing mode' without first understanding the actions that the OBSW will take in response—and may have been involved in specifying what those actions should be.

For example, this command may have a range of effects on the spacecraft, including

- Increased power consumption (if additional actuators, sensors are switched on).
- Misalignment of communication antennas/cameras (if spacecraft attitude changes).
- Entry into safety/survival modes (if the Sun cannot be automatically located, e.g. due to eclipse).
- Depletion of on-board propellant (if thrusters are used for attitude control).
- Activation of faulty hardware (if the OBSW selects inappropriate units to perform the action). More capable OBSWs require ever more time and effort (and cost) to develop. As with all automation, increasing complexity introduces its own opportunities for design faults. The OBSW is only ever as good as the testing to which it has been subjected, and the foresight of the engineers who created it. Once flying, the spacecraft operator retains the final responsibility to ensure that the right command is issued at the right time with predicted consequences. The motivation to get it right is not to jeopardize mission success.

Gregory L. Davis, Raphael R. Some and Andrew A. Shapiro

Effective management of technology development is crucial to advancing state of the art of spacecraft engineering that ultimately enables new scientific discovery. Key elements of this management activity include understanding technology life cycles, composing and evaluating a technology portfolio, encouraging and managing innovation, and knowledge of best practices for technology task management. A technology manager needs to be aware of these essentials within the context of the organizational setting to provide for optimal implementation and results. This chapter summarizes the salient points for each of these key elements, with the intent of providing a useful set of guidelines for the prospective aerospace technology manager.

A useful tool that will be referenced throughout this chapter for evaluating a technology's maturity and monitoring its development progress is the technology readiness level (TRL) scale, introduced in Sect. 2.3.3. After an earlier period of gestation within NASA and other parts of the US government, the TRL scale was formalized in a white paper by Mankins [1] in 1995, and is now widely used across industry, government, and academia to gauge technology maturity. (There are other closely related definitions for TRL [2–4] but the authors will use the NASA definitions.) Technology readiness levels, along with key discriminators and exit criteria, are summarized in Table 21.1 and Fig. 2.10. These criteria are extremely useful in evaluating a

technology portfolio, which is fundamental to the practice of technology management.

The technology portfolio is a set of strategic development activities that, upon fruition, provide an organization with unique technical capabilities relevant to its intermediate-to-long-term goals. Key considerations for managing a technology portfolio include balancing risk versus reward, distributing TRL content, apportioning 'push versus pull' technologies, deciding whether to 'make or buy', monitoring development progress, and ultimately determining return on investment (ROI). In addition to the elements mentioned above, this chapter will also summarize the salient points for each of these considerations, with the intent of providing another useful set of guidelines for the prospective technology portfolio manager.

21.1 Technology and Product Life Cycles

Technologies and products alike have a natural life cycle, ranging from initial concept and technology validation, through technology maturation and initial product development, to product maturation, and ultimately to obsolescence and end-of-life. This dual life cycle is summarized in the so-called Whale Chart depicted in Fig. 21.1, which shows the relationship of technology utility and/or product sales as a function of time. The period between initial product development and end of life is often lengthy and may include development of a line of products differentiated by advancements in the technology, customer communities, quality and reliability levels, and price points, or stated differently, market refinement and segmentation.

Figure 21.1 implies that technology/product development progresses uniformly, beginning with initial technology conceptualization and culminating in a fully matured product line before lapsing into obsolescence. The reality is more chaotic, often characterized by skipping steps or

G. L. Davis (✉)
Mechanical Systems Division, Jet Propulsion Laboratory (JPL),
California Institute of Technology, Pasadena, CA, USA
e-mail: gregory.l.davis@jpl.nasa.gov

R. R. Some
Autonomous Systems Division Jet Propulsion Laboratory (JPL),
California Institute of Technology, Pasadena, CA, USA

A. A. Shapiro
Early Stage Innovation Office, Jet Propulsion Laboratory (JPL),
California Institute of Technology, Pasadena, CA, USA

Table 21.1 Technology readiness levels (TRL) summary

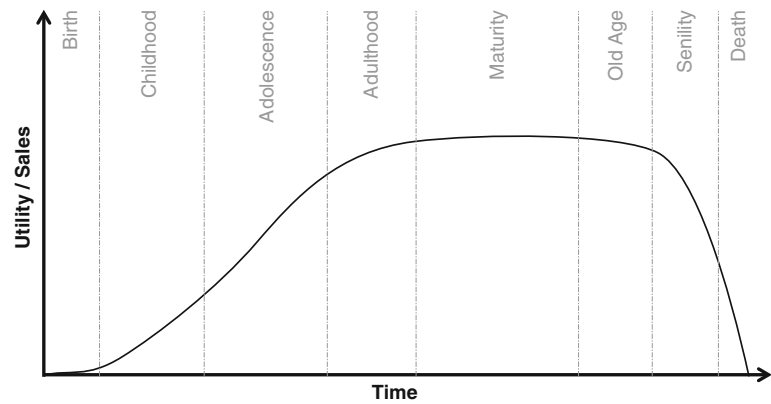
TRL	Description	Incremental Change	Exit Criteria
1	Basic principles observed and reported	Research is <i>transitioned</i> from <i>pure</i> to applied	Peer reviewed publication of research underlying the proposed concept / application
2	Technology concept and/or application formulated	<i>Applications</i> are identified; <i>invention</i> begins	Documented description of the application/concept that addresses feasibility and benefit.
3	Analytical and experimental critical function and/or characteristic proof-of-concept	R&D leading to <i>proof-of-concept validation</i> is initiated	Documented analytical / experimental results validating predictions of key parameters
4	Component/subsystem validation in laboratory environment	<i>Elements prototyped and tested</i> in the laboratory	Documented test performance demonstrating agreement with analytical predictions. Documented definition of relevant environment.
5	System/subsystem/component and/or breadboard validation in relevant environment	<i>Prototyped elements integrated and tested in a space-like environment</i>	Documented test Performance demonstrating agreement with analytical predictions. Documented definition of scaling requirements.
6	System/subsystem model or prototype demonstration in a relevant environment (ground or space)	<i>Representative (engineering) model of integrated system fully demonstrated in a space-like environment</i>	Documented test Performance demonstrating agreement with analytical predictions.
7	System prototype demonstration in an operational environment	An integrated prototype is <i>demonstrated in a space environment</i>	Documented test Performance demonstrating agreement with analytical predictions
8	Actual system completed and “flight qualified” through test and demonstration in an operational environment	<i>Verification and validation completed for the integrated flight system</i>	Documented test performance verifying analytical predictions.
9	Actual system “flight proven” through successful mission operations	<i>Heritage established</i>	Documented mission operational results.

looping back to an earlier stage in the process in order to accommodate the realities of technical setbacks, market forces, and organizational imperatives. Management of this process requires flexibility and understanding of the fact that an orderly progression of technology and product development is a worthy ideal, but pragmatically unrealistic. It is not unusual, for example, for a development team that believes it is at TRL 5 to discover an issue that requires a return to TRL 3 to investigate an anomaly or to develop some aspect of the technology that was not previously understood or even recognized. In the heat of a project, with schedules and budgets looming, this situation can be perceived as a major disruption and a desire to forge ahead regardless may prevail. This approach is usually a mistake, often leading to further schedule delays, budget overruns, and in some cases, failed projects. The successful technology manager, on the other hand, remains focused on the larger objectives and accommodates the inevitable setbacks, resetting the schedule to an earlier TRL if necessary in order to make faster progress later on. Technology managers should expect these kinds of developments and plan for them, building in schedule margin and budget reserves. The

authors typically assume at least one return to a previous TRL level in the TRL 3–5 time frame, and often—depending on the complexity of the technology development—more than one such retrenchment. Similarly, in transitioning from technology to product, where new requirements such as design for testability and manufacturability come into play, it is wise to assume that a significant redesign will be required.

In addition to the technical challenges outlined above, the prospective aerospace technology manager should also be keenly aware of several important industry-specific programmatic factors that are challenges to ultimate success. First, product lifetimes are often measured in decades, not years. Unlike the fast paced world of consumer electronics, for example, where product lifetimes are often 1–3 years and a product line may last for as little as 5–7 years before being superseded by a completely new technology, aerospace products and their underlying technologies are often maintained for 20–30 years or even longer (consider the Boeing 747 product line, now entering its 5th decade of service). Furthermore, aerospace’s long development times tend to require larger capital outlays,

Fig. 21.1 Life cycles as depicted in a whale chart



which must compete internally with other precious research and development (R&D) monies. These long-term time horizons require correspondingly long-term commitments on the parts of the technologist to stay the course in the face of the inevitable technical and programmatic challenges that will arise, and of the technologist and the user to commit to cultivating an extended partnership with each other. Second, many aerospace organizations are culturally risk-averse, engendering conservatism in the adoption of new technology. It is incumbent on the technology manager to recognize and address this conservatism head on. To be successful, both management and potential customers must be presented with a picture showing how adoption of the new technology provides an overwhelming (or enabling) advantage not only in the near term, but for the long term as well. Issues to be confronted typically include not only performance, but also reliability, scalability, flexibility, and adaptability. Thinking ahead and steering the development so that the barriers to adoption are minimized will help to ensure the technology's acceptance and infusion.

The following discussion will distinguish between technology and product development activities—while often intertwined and overlapping, they refer to separate and complementary pursuits. Technology development is the linkage of underlying scientific principles, usually based in physics, chemistry, mathematics, and materials science, to an engineering application upon which future product development is based. These basic principles are often embodied in a model that can be used to predict the behavior of products based on this technology. More specifically, a technology model, which is incrementally developed and verified through the early-to-mid-TRL levels, enables accurate prediction of the behavior of an article that has been designed within the space over which that model has been validated. Products, on the other hand, are articles that have been designed, fabricated, and verified to perform as predicted by the corresponding technology model. In many cases, an initial product development is

intimately intertwined with an initial technology development; however, a follow-on product development may include the need for additional technology development or extension of the space over which the technology model was initially validated.

21.1.1 Initial Technology Development and Product Breadboarding

The progression from TRL 2 through to TRL 3 is often referred to as initial or early stage technology development. Its purpose is preliminary evaluation of the technology or product concept and initial development of the underlying technology model. This early phase of technology development is usually not expensive and consequently is more readily funded; thus it is customary to fund a variety of these low TRL efforts and to weed out those that are less promising before moving on to the more costly-mid-TRL development stage. From a management perspective, this phase is crucial in determining the viability of the technology concept and in gaining an understanding of what will be required for development into full maturation and eventual productization.

The key to successfully negotiating early TRL technology development is to focus the research team on the key issues and not to become distracted with ancillary concerns that can be deferred. While some understanding of the ultimate application of the technology and customer desires can be useful, technology developers at this early stage should not be burdened with achieving an understanding of customer needs. Indeed, customer input at this point can be detrimental to a successful technology development by artificially constraining technologists and inhibiting their creative thinking. Instead, the team should be focused on the key capability desired, stated in relatively broad terms, allowing freedom to explore the technology to determine capabilities, constraints, and dependencies. Extreme rigor in

documentation and testing is not yet required—there simply isn't sufficient time or money for that at this stage of the development cycle.

For early stage TRL development, it is often helpful to enlist participation from universities or other research labs when competitive concerns are not present. Adding outside researchers stimulates innovation, further develops a cadre of technologists for maturation of the technology (as well as early stage development of future related technologies), and may help to control costs. With universities, it is important to realize that their interests lie principally in performing fundamental research and publishing, and that they should not be relied on to 'deliver' a 'product' in the way that an industrial organization would. Instead, when working with universities it is generally best to be able to provide sufficient funding for at least one Ph.D. student and the supervising professor's summer salary over a nominal 3 year period. It is also advisable to require at least semi-annual progress reports, copies of all publishable papers and theses, and any experimental results and models. Arrangements involving multiple universities in related research areas will work best if their research areas are carefully designed not to directly compete with each other, and if they are encouraged to work cooperatively. Holding a semi-annual technology interchange meeting (TIM) is also advisable to encourage cross-fertilization of ideas and to further cooperation. As an example, on a large technology development program one of the authors enlisted 10 universities to work on various aspects of a research problem. Each university was contracted to work in a different area of research, but each of those research areas was a topic in a larger technology development. The contracts provided support for 2–3 Ph.D. students for up to 3 years. Semi-annual TIMs were held to update the universities on overall progress, including the research being conducted by the authors' organization. Several of the universities and professors involved ultimately developed ongoing relationships with each other and with the author. The authors' organization, however, was responsible for extracting useful research information out of the researchers' reports and incorporating these advances into the project. Ultimately, in addition to supporting the authors' technology development, several Ph.D.s were graduated, seminal work was done in one new area of research, and a pipeline of Masters and Ph.D. level engineers from those universities to the authors' organization was enabled.

At the end of this early TRL phase of development, the team should have a 'breadboard', i.e., an improvised test article that has been shown to provide the desired capability in a stand-alone mode under ideal or ambient conditions. By this point, the technology model should be capable of predicting the performance of a similar test article under similar operating conditions. Key parameters should have

been identified and verified. The following management activities are appropriate upon completion of the early TRL work

1. Demonstrating the capabilities of the new technology to upper management as well as to potential users.
2. Projecting the types of capabilities and products that will eventually be enabled.
3. Planning for the next stage of technology development including realistic assessment of cost and schedule.
4. Delineating key risks and forming a strategy to address those risk areas as early in the process as possible.
5. Defining decision points, those points at which crucial decisions regarding the path the development task will take are made, including
 - Alternative technology options
 - De-scoping of activities that cannot be afforded or that do not look promising
 - Termination criteria, those conditions under which it is determined that the technology will not be successful.

These decisions and the criteria by which they are made must be defined early; once development is underway, it is exceedingly difficult to decide to end it. The authors strongly advise following the guidelines outlined above to provide some discipline to this rather loose phase of the development activity, and to help ensure that precious resources are wisely allocated.

21.1.2 The 'Valley of Death'

The progression of TRL 4 through TRL 6 is often referred to as the 'valley of death' because of the traditional difficulty in obtaining the resources required to bridge the technology development from breadboard to prototype in a relevant environment; see Table 21.1. In this evolution, the technology is fully developed and validated through a succession of experimental builds, tests, and evaluations. Typically, the process starts with early breadboard testing in isolation, then proceeds to testing successively refined breadboards in a system environment, and finishes with building a true prototype that is compatible in form, fit, and function with an inertial product. This phase ends when the prototype is tested in the relevant environment; i.e. an environment that is stressing to the technology and representative of one that the initial product offering will experience, thereby validating that the underlying technology will be capable of yielding useful product.

Early stage TRL development begins with an open, relatively unstructured approach and with the team quickly forging ahead to determine viability of the technology. At the conclusion of this stage, both the team and management are convinced that the technology is feasible, and management provides tentative approval and sufficient funding

to proceed to the next level. Over the next months and years, the team will need to transition from rapid early technology development to a more deliberate development methodology for the mid TRL development to be successful. In particular, specifications are now written, designs are documented, experiments and tests are designed and documented more thoroughly, testing becomes as important as designing the test article, system considerations become important, and issues of testability and manufacturability begin to arise. This transition will not occur suddenly: rather this evolution needs to be tempered to ensure that money and time will not be wasted on the finer points of producing a product too early in the development stage, and that development issues are addressed in the appropriate order as planned at the end of the early TRL development stage. The technology manager is advised to formally notify the team of these new rules of engagement and to augment the team with the necessary complementary talents and possibly phase out others. A key issue in the transition from early to mid TRL is that the early TRL technologists will chafe at the added responsibility of documentation, test, and analysis; whereas the just-added team members who are accustomed to a more rigorous approach to development—especially if they come from a product development background—will likewise chafe at what they perceive to be a lack of rigor, of playing fast and loose with the technology, and of not understanding the needs of the product developers and customers. The manager's job is to appreciate these differing points of view and to manage the evolution: to continually show both sides the need and advantages of working through these evolutionary stages; to focus on the key issues; to maintain progress in order to maintain funding; and to return often to the big picture, which is needed to keep the team motivated and aware of where it stands in the overall development process.

As development proceeds through TRL 5 and 6, i.e., evaluating the technology in a system, then in a 'relevant environment', and finally to developing a true prototype, there is ever greater need to understand the end use and end-user environment. Obtaining customer input and utilizing experienced aerospace product development engineers to bring real-life experience and understanding of realistic usage conditions is critical to successfully developing the technology and to ensuring successful product development and infusion. Many technology development efforts have failed because the technology development team did not understand either the real needs and priorities of the customer, or what was required to take a technology to product. As an example, several years ago one of the authors witnessed what was portrayed as a highly successful technology development that delivered several working prototype units. Given the capabilities of the units and the benefits

accrued to the system that incorporated them, the developers expected straightforward and immediate adoption of the technology and infusion into the target system. Millions of dollars had been spent over several years to develop this technology, the test articles worked extremely well, and the model performance predictions were accurate. From the developer's perspective the development was highly successful and well worth the money spent; however, the technology was not adopted and the prototypes never made it to product. Management asked the author to investigate the situation to determine what happened, and to ensure that the experience would not be repeated. During the investigation, the product development engineering and customer communities identified the following issues: (1) lack of documentation of the testing and validation activities, (2) difficulty in manufacturing, including the cost of manufacture in addition to assembly and inspection, (3) long term reliability concerns, and (4) perceived fragility of the final product to environments encountered in the system. The author observed that in addition to these explicit issues, an implicit issue was the lack of familiarity with the technology and a lack of confidence stemming from lack of insight into the development process. The author was then asked to restart the technology development activity in an attempt to salvage the previous investment, thereby providing the badly needed capabilities that had spurred the original technology development project. After technically analyzing the prototype and determining its strengths and weaknesses as well as its benefits and cost, the author agreed to lead a new technology development task, 'resetting' the technology maturity level to TRL 3 and planning a new development from that point on. The author then proceeded to engage the organization's lead product and system developers, bringing them onto the team and, with their assistance, helping the technology developers to understand the needs of product developers and customers. The customer community was similarly engaged as the development effort continued to ensure understanding, familiarity, and acceptance of the technology, as well as to ferret out adoption issues early in the process.

21.1.3 Productization and Technology Infusion

The progression from TRL 7 through to TRL 9 is often referred to as the region of productization and infusion: the development of an initial product, typically quite similar if not identical to the prototype, and the infusion (or sale or insertion) of the product into a system. Arriving at this stage means that the 'valley of death' has been successfully bridged into the region of product development. The

Table 21.2 The growth path matrix. Image: Noel Capon, reprinted by permission [5, 6]

		→			
					↑ Risk
Market	New	Market Growth 2: Market Expansion	Business Expansion: Concentric Products	Conglomeration	
	Related	Market Growth 1: Market Extension	Business Extension	Business Expansion: Concentric Markets	
	Existing	Market Penetration	Product Growth 1: Product Extension	Product Growth 2: Product Expansion	
		Existing	Related	New (to firm)	
		Product or Technology			

ultimate customer now becomes the primary focus, and attention turns to satisfying the manager who decides to buy into this product and stake his reputation on its long-term viability.

As mentioned earlier, the aerospace customer isn't making a decision on a product with a 3 year lifetime; the customer is making what he hopes will be a 20 year commitment to a technology, a product, and a product line. Cost, price, affordability, and profit are extremely important considerations in this stage. The ultimate drivers for success will be benefit to the customer and ROI for the product developer/manufacturer. At this point, the product development manager's job is to elicit from the marketing and sales personnel, as well as from potential customers, the needs and desires of the broad customer community, and to extract from those requirements and wishes the key features that will be engineered into a product line plan. It is also the responsibility of this manager to continue to refine the product to maximize customer benefit, thus ensuring continued customer loyalty and product marketability while concurrently maximizing profitability of the product and the resulting ROI to the organization.

From a management perspective, the technology development is complete and product development—a very different task, that deals with manufacturing engineers, customer representatives, marketing and sales personnel, product line managers, parts specialists, and supply line managers—is now underway. This environment is one that many technology developers, including technology managers, are ill equipped to deal with. Successfully adapting to this environment, however, will ensure the long-term success of the technology, the profitability of the product line, and the confidence of upper management in the technology manager's own abilities.

Infusion of the first product is absolutely critical: success or failure of the technology and its future development is determined at this point. Typically, product engineers rather than technologists implement productization. Technologists may be retained on a consulting basis through its initial stages to address transitioning issues—and it is incumbent on management to ensure that they be made available as needed—but it is not necessary that they be dedicated to the

productization effort. Typically, the technologists should be moved on to the next project, continuing development of this technology area.

The productization process is characterized by a high degree of rigor in design, documentation, testing, and qualification by those who best understand how the product will be deployed in a system. Afterwards, a transition to production manufacturing ensues, requiring development of manufacturing tools, documentation, and methods, all specialties of product design and manufacturing engineers. Design for manufacturability and testability comes to the fore at this point, and if these factors have not been previously considered, they will often necessitate a redesign of the prototype. Similarly, as the product is aimed at insertion into different target systems, features will often need to be modified or added to meet customer needs or concerns. The product development manager must develop a longer term plan so that engineering time is not wasted and potential efficiencies are not lost in the rush to get the product to market, while at the same time keeping the engineering team focused on getting this first product qualified and shipped. Some years ago, one of the authors was involved in developing a highly specialized product for a specific customer. In looking at the overall market, however, it was clear that a similar capability would be of value to a broader market place. The product was designed to meet the immediate customer's needs in a timely fashion, but additional features were also incorporated to facilitate easy adaptation of the product to other customers. The result was a profitable product line that was easily adapted without costly redesign for a broader market over a several year period, and one that provided each customer with a unit 'customized' to his specific need at a low cost. Because these customization features did not significantly increase product complexity, the first product insertion could be done within the schedule and budget allotted, and without appreciable increase in longer term manufacturing cost.

Once the first product has reached some measure of success, development of a product line may be possible, providing differentiated products for different market segments or increasingly more advanced products for the same market segment; see Table 21.2. In many cases, advances in

technology will benefit the customer by allowing either a reduction in cost or an increase in new features at similar cost, thereby increasing the utility and profitability of the product line. In developing a product line, it will be useful to leverage a core design and implementation approach across all members of the product family by defining a baseline—typically lowest price—product. Where possible, features that can be incorporated at no additional cost can be used to improve the product or to tailor a version of it to a specific market, possibly at a higher price point. As the market matures, it will often be possible to offer the same product(s) screened to different quality levels, with corresponding differentiation in price.

21.1.4 Product Obsolescence

Towards the end of the normal product/technology life cycle (Fig. 21.1) newer technologies will inevitably overcome the capabilities of the existing product line. This situation is not, however, the end of the story. The existing product line must be maintained indefinitely due to legacy products that have been fielded—in the aerospace market, a manufacturer cannot simply abandon the customer or the product line. Meanwhile, new technologies should be adapted to fill the niche of the old products: those with better performance; lower mass, power, volume, and cost penalties; or those with higher reliability. These new technologies should be made ready to begin their life cycle as older technologies reach the end of their life cycle. And as the organization continues to innovate, the cycle continues.

21.2 The Strategic Technology Portfolio

As mentioned previously, a technology portfolio is a set of strategic development activities that, upon fruition, provide an organization with technologically unique capabilities relevant to its intermediate-to-long-term goals. To develop a strategic technology portfolio, it is necessary first to have a concept of what is strategic for the institution. In general, this requires strategic planning, a process often ignored in many institutions. In both for-profit and not-for-profit institutions, there is significant competition for resources, and the immediate problem or ‘this quarter’s numbers’ often take precedence. An executive assigned to manage a technology portfolio must first establish the importance of strategic planning and investment as an institutional priority, and second, continually remind management—both middle and upper—of its importance. It is critical that the responsible manager firmly establish the importance of maintaining a critical threshold level of resources for an effective and consistent strategic management program,

including developing and periodically maintaining a ‘strategic technical plan’ or ‘roadmap’, and monitoring the portfolio. Regardless of the industry, nearly all technology investments need to clearly be tied to planned products as defined in the strategic technical plan.

It is difficult to go somewhere—or to even to know that the destination has been reached—without a plan or map of some kind. Technology plans or roadmaps can take many forms, and their particular details are best left to the organization. There are, however, some common features that should be present in nearly all forms of these documents. At a minimum, the management team needs to agree on and to codify, at least at top level, what business it is in. In particular, the products (or product lines) and services the organization provides as a business need to be defined. This activity may seem obvious; however, there are many instances where this assessment has not been performed. Organizations have often been surprised to find a disparity between their actual and perceived business activities. Once particular products have been determined, their future *strategic* endpoints need to be identified, along with candidate tactical paths for arriving there. During this process, the following types of questions should be answered

- What are the key technical strengths?
- What are the technical weaknesses?
- Are the technical strengths in line with the future vision?
- Do the product lines need to grow [5]?
- What is the market expansion strategy, i.e., which growth path to follow [6]? (see Table 21.2)

One outcome from answering these questions should be a realization by the organization of which technologies should be developed internally, leveraging institutional strengths in order to maintain a competitive advantage versus which technologies should be procured from other sources, acknowledging institutional weaknesses and ceding the advantage to the competition. These so-called ‘make or buy’ strategies should be addressed in the strategic plan.

An area that is often overlooked is a strategy for re-deploying technical personnel into growth areas of the current portfolio and away from non-growth areas or those no longer of interest. It should be noted that occasionally technical specialists can be moved from one discipline to another. Unlike many other business personnel however, technical personnel are not fungible; senior technical personnel tend to have an extensive historical knowledge base that would take even a highly competent specialist from another field many years to fully learn.

In addition to securing the importance of and resources for technology, defining products, defining technical directions (paths for infusion), assessing strengths, evaluating technical personnel, and planning growth, a solid strategic technical plan should also address strategic alliances, university relationships, communications and outreach

(conference and committee participation, etc.), publication strategies, and patent/proprietary information (company secret) strategies.

Regardless of the source of a potential technology innovation, if it is not in the existing strategic plan (and presumably part of the business strategy), the strategic plan is in need of being updated. If neither of these outcomes makes sense in a business context, the investment should not be made.

21.3 Risk Versus Reward

A fundamental consideration in assembling a technology portfolio is assessing the balance of risk versus reward. Risk, in the context of technology development, represents an assessment of the probability of an adverse outcome set against the consequence of that adverse outcome. One type of risk occurs when an investment in a particular technology does not result in a successful product (or worse results in a successful product for your competitor, e.g., the Apple interface from Xerox PARC). The consequence is typically tied to the magnitude of the investment: if the investment is a minor percentage of the institution's profits (or budget for non-profit institutions), then the consequence of an adverse outcome is low. If, however, the investment is the entire profit of the company for several years, then the consequence of an adverse outcome would likely be catastrophic. A related type of risk is overweighting the portfolio in a particular technology at the expense of others. If the company has a number of candidate technology investments, they all should be represented in the portfolio to the extent that a critical resource mass can be provided. Yet another type of risk that is often overlooked—and one that can be even more critical—is the risk of *not* investing in a new or advancing an existing technology, also known as inverse risk. A striking example of this is the case of Kodak, an organization that nominally invented the digital camera but has now filed for bankruptcy and may go out of business at a time when digital cameras have taken over the market and are highly profitable. Kodak made an excellent initial investment in developing the digital camera; however, it failed to invest in the infusion of this technology, with catastrophic consequences. As extensive discussion of inverse risk may be found in [7].

Reward, in the context of technology development, represents an enhanced capability, product line, and/or income stream resulting from an initial investment. The most common measure of success, as viewed by most managers, is the monetary value of the product sales resulting from the technology infused into existing or new

products. This tangible benefit is straightforward to calculate using standard techniques from finance.

An excellent example of a critical, successful risk-reward trade was the commercial development of the Boeing 747 aircraft. In the mid-1960s, with commercial long-distance air travel starting to become widespread, Boeing anticipated the need for a 'jumbo jet' to replace its aging 707 product line. There were huge partner, technological, and financial risks associated with such a development: organizationally, Boeing relied on a partner, Pratt and Whitney to develop the new engine; technically, an entirely new high-bypass jet engine was required, as was a massive new assembly facility to construct the aircraft; and financially, the company had to leverage itself heavily to complete the development program. This bold development risk has been richly rewarded since the rollout of the first 747 in 1970, with Boeing holding a monopoly in very large aircraft production for many years afterwards, and with this product line remaining alive and well today as it enters its 5th decade of service with over 1,400 deliveries.

In addition to tangible rewards, there are ancillary and sometimes hidden or intangible benefits from technology development projects that may accrue regardless of the level of financial success. These intangible benefits, which can be quite valuable to the organization's future strategic position, include developing a familiarity with the technology area, training personnel, enhancing organizational capabilities, establishing links to academia and other research organizations that can further future technology objectives, and establishing a reputation within the industry for promoting technological advancement and for expertise in the field. As an example, several years ago one of the authors was involved in the development of a new type of computing system for advanced spacecraft. Although good progress was being made, several years into the effort the climate had changed and the project objectives were no longer deemed to be of high value to the organization. After having spent tens of millions of dollars on this development, the project was rather abruptly terminated, the team disbanded, and the result was deemed a failure for not having achieved its stated objectives. That outcome was, to say the least, somewhat disheartening. Within the following 3 years, however, drawing on expertise developed during the terminated project, a new project with similar goals was initiated. With reduced funding, but with a head start due to experience gained in the previous 'failed project', the new project was able to accomplish its goals quite handily. Furthermore, the work done on the original project resulted in strong participation in technology conferences, published papers, and establishing the organization as a leader in the field, a reputation that was of significant benefit in the

marketing and infusion of related products from the new project.

21.4 Push Versus Pull Technologies

Another important consideration in assembling a technology portfolio is apportioning the mix of ‘push’ versus ‘pull’ technologies to diversify risk. Pull or ‘top-down’ technologies are those that senior management has identified as being strategically important to the organization, and those that may provide a competitive advantage when infused into a product line. Identifying these technologies provides institutional focus, and a rationale for apportioning scarce resources in what is always an oversubscribed R&D budget. On the other hand, ‘push’ or ‘bottom-up’ technologies are those that technologists themselves identify as being potentially valuable to the institution. They provide a counterpoint to the ‘pull’ technology process: although management may try to select the most promising technologies, frankly it isn’t smart enough—and never will be, no matter how talented—to always pick the winners. Allowing for push technology developments allows for the unanticipated breakthrough development (e.g., 3 M Post-it[®] Notes) while simultaneously encouraging the creative and entrepreneurial behaviors so important to innovation.

21.5 Optimal Portfolio Composition

The optimal technology portfolio should be balanced across many dimensions: as mentioned earlier key criteria include balancing risk versus reward, distributing TRL content, apportioning ‘push versus pull’ technologies, deciding which technologies to include or exclude (whether to ‘make or buy’), monitoring development progress, and ultimately determining ROI.

Several criteria for managing technology investments have already been mentioned. To recap, first a prudent risk posture vis-à-vis an organization’s resources should be adopted; only in the direst of competitive settings should a ‘bet the company’ approach be pursued. As a corollary, investments should be spread across candidate technologies to the extent that critical resources are available. Second, through strategic planning, an organization should decide which technologies to include, and which to exclude (‘make or buy’) in its investment portfolio. Last, an appropriate mix of ‘push’ versus ‘pull’ technologies should be present; although there are no hard and fast rules for this ratio, experience has shown that a portfolio composed of

approximately 66–75 % for ‘pull’ and 25–33 % for ‘push’ technologies is most effective.

An ideal technology portfolio will also span the complete spectrum of TRLs. The total investment in each range (low 1–3, mid 4–6, high 7–9) will vary substantially depending upon on the individual company’s status and projected needs. One reasonable strategy is to allocate the same investment capital for each range. Because lower TRL investments are often theoretical or crude in nature, they typically will cost substantially less to demonstrate than more mature technologies. Consequently, an even distribution of funds may result in ten low-TRL projects, three mid-TRL projects, and a single high-TRL project. A natural down-selection of concepts as one proceeds up the TRL ladder occurs because many interesting theoretical or low-TRL concepts prove to be impractical or too costly to move forward. Regardless of the maturity of the technology, every technology should have a clear, well-defined path to infusion into a real product or it should not be in the portfolio. This infusion path should be revisited at each new TRL step. A technology portfolio manager must accept the idea that most technology concepts will fail and never become infused. If one in ten technologies progress from TRL 1 to TRL 9, the investment program should be considered an outstanding success. Most technology portfolio track records are closer to one in a hundred.

The effective manager will strive to incorporate all of these considerations into a balanced technology portfolio. Creating such a portfolio is not a purely mechanical exercise, it must also be coupled with good judgment—often nuanced—to provide the organization with a technology investment program with the greatest probability for product infusions. Once composed, the portfolio must be monitored on a regular basis and its returns measured in order to adjust the investment mix.

21.6 Return Metrics

A simple axiom nearly always holds true: whatever gets measured will improve; judicious selection of what is measured will inevitably lead to better performance. That said, ROI metrics for technology are notoriously difficult to implement, but critical nonetheless for providing rationales for disciplined decision-making and resource allocation. A common approach is to use techniques from classical finance theory. The cost to mature the technology from concept to infusion can be reconciled against the future income stream of the final product (both factoring in the time-value of money) to give a total ROI. In a variation,

leveraged return arises when an investment in a specific technology enables the opening of a new market; the profit from the entire new market—even though the technology may not have been used in all of its products—can be used to compute the ROI. Another financial method is to calculate return on assets (ROA) or return on net assets (RONA), thereby factoring in laboratory and capital equipment costs (which may be difficult to compute if multiple projects use the same lab and equipment) to give a measure of how efficiently the organization utilizes its assets. In general, using ROA/RONA for technology investment decisions appears to be of diminishingly small value; ROI, however, appears to be more useful. In addition to ROI based purely on profit, some investments' returns are primarily from intellectual property. A well-developed patent portfolio, with its attendant rights to exclusivity and/or licensing/royalty income streams, can often be the most profitable product of a company. Yet another metric for success is how many technologies advance at least one TRL level in a given review timeframe. This can be determined by periodic peer review with a 'gate' process using predetermined exit criteria similar to those depicted in the far right column of Table 21.1. There are many other metrics that can be used, and individual companies need to decide what works best for them. Regardless of the approach used, some form of ROI should be calculated for product investments, even for the lowest TRL levels.

As mentioned previously, in addition to tangible financial benefits there are often intangible benefits from technology investments that are difficult to quantify, but are extremely important nonetheless. One such benefit is increased name or brand recognition, even if the technology investment itself is not extremely profitable. A good example of this is the benefit derived from the technology investment in the Mars Pathfinder airbag landing system (ALS). Although this investment was not exceptionally profitable for the medium-sized firm that developed and built the ALS, the firm received high profile accolades for the system's successful performance, gained substantially increased name recognition, and ultimately gained better access to space and other related markets. The rewards of technological leadership derived from participating in activities such as published peer reviewed papers, conference papers and presentations, professional committee participation, and hosting of workshops or forums can also be of high intangible value. In these cases, the return is in the technical community's recognition of leadership: the value is, presumably, an increased and more loyal customer base from those customers who want to be perceived as being at the leading edge, or those customers who perceive the organization's value in technical leadership. Some customers will pay a premium for this perceived value.

21.7 Managing Innovation

Innovation, on its own, is of little or no use; it is the development of innovation into superior products that is of value. Despite many opinions to the contrary, innovation can be systematically managed [7]. Virtually every engineer, technologist or scientist has at least one great, innovative idea; what will make the idea develop into a successful product has much to do with how these concepts are managed and how well disciplined the researcher is. The term researcher is used because engineers, scientists, and technologists can—and should—perform research into innovative concepts encountered while performing their regular jobs. The ensuing discussion is divided into outlining researcher and management responsibilities under the innovation process and afterwards, describing types of innovation.

21.7.1 Responsibilities in Managing Innovation

The researcher who has an innovative concept needs to surmount several hurdles to be successful. First, and most importantly, the researcher needs to understand exactly how his innovation will benefit his institution's current or future products. If the ultimate instantiation of the innovation will not benefit the organization, then the researcher has several choices: drop the concept and work on something else; or with management concurrence, pursue the concept independently within the organization; or leave the organization to pursue the concept in another company. Assuming that the innovative concept has potential benefit to the institution, then the researcher has to surmount two additional hurdles: securing funding, and defining a pathway from the innovation into an initial flight product.

The researcher has a responsibility to persuasively answer a series of questions to convince management or a proposal sponsor to give the innovative concept funding. The following is a version of these so-called 'Heilmeier questions', first posed by George Heilmeier (born 1936) of DARPA in the 1970s [8].

Questions to be answered for a successful innovation development.

1. What are you trying to do? Articulate your objectives using absolutely no jargon. What mission need are you trying to meet?
2. How is it done today, and what are the limits of current practice? What is the current state of the art? Describe the limitation of the state-of-the-art to meet the need.
3. What's new in your approach and why do you think it will be successful? Why does the current technology not

meet the mission need? What technology needs to be developed to meet that need?

4. Who cares? If you are successful, what difference will it make? What are the unique benefits to the mission? Why do you think you can beat the competition?
5. What are the risks and the payoffs?
6. What is your unique innovation to meet the need with respect to the competition?
7. How much will it cost? How long will it take? What are the major deliverables? Why is there a need to do this work?
8. What are the mid-term and final 'exams' to check for success?
9. Why should the sponsor fund this?

By answering these questions thoughtfully, before any significant effort is invested, the innovation has a much greater chance of becoming useful. By the end of this process, there should be a clear indication whether the innovation has a chance for success or not. If so, funding should be sought through many avenues; if not, no additional effort should be spent on the concept. It should also be noted that, over the development cycle, answers to many of these questions are likely to change; this should not be viewed as a barrier to either the researcher or to management, but rather as an opportunity to realistically reassess the innovation's prospects for ultimate success. Innovation, by nature, is dynamic and should be expected to change.

Embedded in question 7 above are some of the more critical issues for the researcher to focus on. Formulation of a development strategy with key milestones and deliverables—these need not be extensive or numerous—will force the researcher to plan extensively, probably the most important thing that can be done to ensure an innovation's successful realization. One tool that can be useful in this regard is some type of project planning software. Estimating the cost to achieve each of the milestones and deliverables, including capital equipment and materials, is time well spent. It should be noted that these plans will inevitably change (and will need to be revised); however, the planning exercise focuses thought and facilitates accommodating changes. Having a well thought out plan and cost estimate will make obtaining funding from any source (internal or external) much more likely.

With a plan and funding, the last remaining activity is to effectively track progress to determine whether the innovation development is on plan, off plan, or that the plan is in need of revision. Periodic reviews by management are a useful tool for enforcing this discipline. Reviews should be held at least biannually, and quarterly reviews are usually beneficial. Concerns are often heard about too much oversight and too many reviews stifling innovation. In the authors' experience of managing innovation over several

decades, this has not been found to be true; in fact, requiring a small amount of discipline from researchers tends to focus them rather than stifle them, and a lack of discipline in forming a plan often results in wasted resources and no advancement of the innovation. Of course, management can go overboard by requiring onerous reporting and reviews, and obviously this situation must be avoided.

Institutional management has a responsibility to create an environment that fosters and stimulates innovation. There may be some arenas in which innovation is less critical, such as in safety procedures and production lines, but it is absolutely critical for technology development. Prominent ways that management can create an environment to encourage innovation are given below (there are many more)

1. *Buy-in*—Demonstrate by actions and words that innovation is highly valued as an institutional priority. Innovation must be recognized and rewarded, and its progress included in annual performance reviews.
2. *Generic resources*—Provide a small amount of funding for all technical staff to 'fool around' with creative ideas. Some companies fund 10 % or more of their technical personnel's time in this way.
3. *Dedicated resources*—Reserve dedicated internal research and development (IR&D) funding to develop the best concepts. One strategy is to have several levels of funding: one level of slightly larger funds to develop more mature concepts, one of smaller funds for less well developed concepts, and one for workshops where field experts are called into help clarify and focus a concept.
4. *Periodic solicitations*—Issue periodic calls for new concepts. Review many and select a few for development funding. Again, consider several levels of funding.
5. *Proposal support*—Provide help and resources for proposals from external sponsors (NASA, DoD, DoE, DARPA, etc.). Provide a bid and proposal (B&P) budget to enable high quality proposals. Provide proposal assistance and independent peer review.
6. *Teaming*—Encourage teaming within and external to the institution. Provide a teaming budget. Reward individuals that successfully team with outside institutions; reward managers that arrange internal teaming arrangements. Cultivate long-term relationships with outside entities at all levels including executive management.
7. *Academia*—Provide for a university-funding program. Build relationships with top universities in the field and fund them on a regular basis. Remembering that it takes 5–6 years for a typical Ph.D. student to graduate, provide longer duration funding.
8. *Planning and documentation*—Require documentation and planning for concepts that people want to develop. Have them answer Heilmeier's questions.

9. *Progress reports*—Have periodic reviews of progress (short time per task). Check plans against milestones and deliverables. Examine spending rates and consider a simple earned value system (see following discussion).

A few additional words should be said regarding research task selection and progress reviews. Most aerospace institutions would have an abundance of innovative concepts if the technical workforce were allowed or encouraged to pursue them. Solicitations for innovative concepts should be institution-wide, and all technical personnel from new-hires fresh out of school to the most senior technical staff should be encouraged to respond (if only senior individuals are encouraged, it will indicate that not all innovations are welcome; young researchers may be paired with senior researchers to help balance out ages). For task selection, the institution needs to develop a systematic way to review concepts that should be strictly merit-based. A review board of 5–10 highly experienced individuals from diverse backgrounds, each representing an important area of the institution, typically works well. Evaluation criteria should include

- Innovation—has it been thought of before?
- A clear pathway to a flight application.
- Alignment with institutional goals.
- Technical feasibility.

Tasks should be ranked using these (or similar) criteria, and the number of tasks selected from the top of the list according to the budget available.

With regard to the review process, again the institution should use a diverse board of 5–10 experienced technical individuals, its purpose being to determine whether the assigned funds were being expended wisely. Progress, achievements, milestones, and expenditures should be evaluated. A simple way to do this is to use two ‘earned-value’ indices, there are many possible such indices, but two simple ones that have proven to be effective are (*% schedule/% funding*) and (*% completed milestones/% funding*). The tasks are on plan if both indices are near one.

21.8 Types of Innovations

There is a wealth of literature on this subject, but broadly speaking innovations tend to fall into two categories: incremental and disruptive. Both are important. Incremental innovations, which take current technologies and improve upon them, serve a critical function for remaining competitive in established businesses and product lines. Small increments in infusing new product technologies do accumulate, and more often than not they make the difference in remaining ahead of the competition; furthermore, innovations that are in line with conventional thinking are easier to adapt and are much more likely to be infused quickly. By far, most industrial research as well as most university

research is incremental. Very few individuals or organizations have the mindset to do anything else.

The alternative to incremental innovation is ‘disruptive innovation’, innovations that completely change the landscape of a particular product or product line. Disruptive technologies have some common characteristics [9]: they tend to revolutionize the field they are in; they often provide lower quality or performance but at a greatly reduced cost; and they also tend to be so disruptive that the existing technology becomes irrelevant. Many such examples exist: the internal combustion engine, nylon, the personal computer, the low-cost CCD for cameras, the Sony Walkman, and the Apple iPod. In each of these instances, the resulting performance was inferior (at least initially) but the product was inexpensive and could be made in much higher volumes.

Disruptive innovation is much more difficult to anticipate and to manage than incremental innovation. In terms of management responsibilities, some of the previously listed items will easily work and some will be more difficult; they are re-summarized below.

1. *Buy-in*—Much more difficult, particularly with senior management. Ideas will be perceived as irrelevant to conventional products and may displace existing products in which the management has a vested interest.
 2. *Generic resources*—This approach should still work well.
 3. *Significant resources*—This approach will work well at low TRL, but will meet more resistance as the technology progresses up the TRL ladder. A well-selected review board can help to mitigate this problem.
 4. *Periodic solicitations*—This approach will work well so long as it is made clear that disruptive concepts are allowed.
 5. *Proposals*—This approach remains absolutely necessary, but more difficult to fund. Peer review needs to be open-minded.
 6. *Teaming*—Teaming may be difficult for disruptive concepts because the partners will need to adopt the disruptive vision.
 7. *Academia*—Mixed, as few academicians are working on disruptive frontiers, but their research bias for low TRL can lead to fundamental breakthroughs.
 8. *Planning and documentation*—Still necessary and applicable.
 9. *Progress reports*—Still necessary and applicable.
- Obtaining buy-in for disruptive innovations from both the senior management and the technical staff is perhaps the most difficult challenge for the organization. One strategy to address this issue is to train the senior management team to understand and to seek out disruptive technologies; having it completely embrace disruptive concepts would be ideal. Senior management’s receptivity to disruptive innovation should set an example, aiding the technical staff in doing the

same. One hurdle that will be a challenge is for the technical staff to overcome its natural inertia to forced progress; some staff will resist having their previous work made obsolete. Another hurdle is the creativity barrier; it is difficult for many staff to think ‘out-of-the-box’, tending instead to be constrained by what can be accomplished with minimal risk.

If creative concepts can be unleashed without penalty and even encouragement from management, many ideas should then come forward. Once a disruptive concept has been identified, it needs to be supported with the correct level of funding. Excessive funding will attract unwanted attention and a plethora of non-contributing personnel; too little funding will eventually kill the concept by starvation. These respective levels are organization-specific, but a savvy manager will recognize where these levels are for his or her own institution. Once the disruptive concept is identified and funded, programmatic tracking (documentation, planning and progress reporting) become critical to keep the concept on track. Without oversight, the concepts tend to stray into territory that is not compatible with institutional needs. The person who conceives of the disruptive concept is often not the best one to carry it forward and careful team selection, including low-level leaders with a ‘vision’ of where the program needs to go, becomes key. In this situation, the individual conceiving the concept needs to be well rewarded in order to encourage the generation of more disruptive concepts. This is particularly true when another person is selected to carry the new concept forward.

21.9 Reward Systems

It is axiomatic that desired behaviors should be rewarded behaviors. Innovation is no exception; in fact, effective reward systems are particularly important in this arena. In the example immediately above of a disruptive innovation, clearly there needs to be a significant award for the initiator, even if that person is not selected to carry the concept forward. Rewards for innovators can be provided in a variety of ways, including

- Management and peer recognition
- One-time monetary awards (bonuses)
- A percentage of return from the product
- Salary increases
- Promotions
- Choice in future assignments
- Paid sabbaticals
- Technology promotional paths equivalent to those for upper management.

A few additional words are in order regarding reward systems. An excellent motivator for providing a clear pathway to product infusion—always very challenging—is to provide a percentage of return on a fielded product directly to the innovator for his or her invention. And finally, the last reward mechanism in the list is probably the most important. Sadly, it is a well-known phenomenon that many promising young technical innovators jump over to management when approaching mid-career because the financial and career rewards for a technical career path are inferior to those for the management track. Technologists need to have the freedom to be creative and to move up in an organization through their contributions as successful innovators, and be appropriately rewarded for doing so.

21.10 Organizational Implementations

In the context of this chapter, the primary function of the organization is, through innovation, to transition technology from concept to product (market). Classic organizational behavior texts [10, 11] consider the organization to have multiple facets, e.g., structural, political, and cultural; this section will consider only the structural aspect. Traditionally, aerospace technology-oriented organizations are structured along functional, product, or matrix lines; a discussion of other organizational forms may be found in [11].

21.10.1 The Functional Organization

The traditional functional (or departmental) organization is the oldest organizational form and is structured along technical specialty or discipline lines. Staffs are grouped by discipline, each arranged in its own hierarchy, and they work independently of other departments or associated projects [11]. Advantages of this approach include encouraging deep levels of technical expertise that foster more intimacy with technologies under development, and having so-called economies of scope that facilitate reallocation of resources across tasks within functions. Disadvantages of this organizational approach include lack of coordination and integration across discipline lines and poor responsiveness to market changes [10].

In a functional organization, the technology manager must be cognizant of these coordination, integration, and market awareness challenges. The organization will tend to obscure these issues, leaving an inexperienced technology ill prepared to deal with larger projects that require interdisciplinary teams, or with changing market conditions that

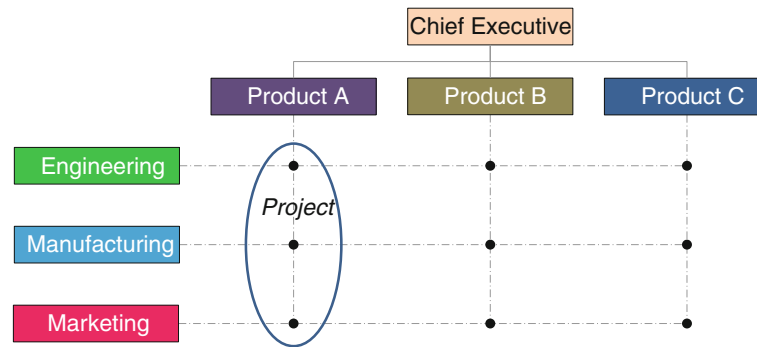


Fig. 21.2 A functional-product matrix organization

render technology and product lines rapidly obsolete. Developing both formal and informal relationships with other functional technology managers as well as with marketing and sales managers, holding regular interchange meetings, and developing other market information sources is a viable strategy that will help the functional manager to remain influential and effective in the face of these forces.

21.10.2 The Product Organization

The traditional product (or project) organization aggregates staff from diverse functions to produce an output or product. Advantages of this approach address the disadvantages of the functional approach; helping to ensure better coordination and integration across functions, increasing responsiveness to market changes, and providing for clear strategic organizational focus. Disadvantages of this organizational approach include a possible erosion of discipline focus to the detriment of new technical discovery, and a focus on existing product lines to the neglect of other possible new opportunities [10].

Technology managers will generally retain a strong connection to their technology discipline, but those who have drifted from technology development into product development may have difficulty keeping up with advances in their field, becoming insulated from new capabilities and applications. Maintaining connections with academic and research institutions, participating in technology conferences, and generally keeping up with advances in their technology area is a straightforward strategy that will offset the relatively minor disadvantages incurred by working with a product focus.

21.10.3 The Matrix Organization

A matrix organization is a hybrid structure that seeks to address the shortcomings of functional and product

organizations by blending the advantages of both into a two-dimensional management structure: one dimension organized by function, the other by product, with the matrixed (or projectized) staff reporting to two managers; see Fig. 21.2. In theory, this structure relies on the competing interests of functional and product management to produce an optimal outcome, i.e., functional management's interest in preserving technical integrity (at the potential expense of getting product to market) is balanced against product management's interest in getting product to market (at the potential expense of technical quality). In practice, this optimal balance is not always achieved, with one dimension of the matrix dominating the other and leading to products that move to market either too slowly (functional dominance) or too quickly (product dominance). A summary of the various organizational outcomes depending upon the relative influence of the constituent parts of the matrix is summarized in Table 21.3 [12]. In recent years, this organizational form has fallen out of favor because of the aforementioned drawbacks and lack of clarity in reporting lines [10].

For larger projects, where it is feasible to do so, matrix organizations will often 'soft projectize', i.e., form project teams whose members are co-located in a specific area dedicated to the project team and somewhat isolated from their functional organization. Once the project has been completed, the team is disbanded and personnel are returned to their home organizations. If managed correctly, with proper attention being paid to rotating individuals between project and functional organizations, this technique can ameliorate some of the drawbacks in matrixed organizations outlined above. More specifically, projects can retain access to dedicated experts as needed, and domain experts can return to their technology organizational roots where they have the opportunity to catch up on advances in their field, perform research and development work, and bring back to the technology organization an understanding of real-world project and product issues. A key challenge for the institution is to manage this personnel rotation, balancing, at a

Table 21.3 Organizational influences on projects; © Project Management Institute, reproduced with the permission of PMI [12]

Project characteristics	Organizational structure				
	Functional	Matrix			Project
		Weak	Balanced	Strong	
Project manager's authority	Little or none	Limited	Low-moderate	Moderate-high	High-total
Resource availability	Little or none	Limited	Low-moderate	Moderate-high	High-total
Budget control	Functional manager	Functional manager	Mixed	Project manager	Project manager
Project manager role	Part-time	Part-time	Full-time	Full-time	Full-time
Project management administrative staff	Part-time	Part-time	Full-time	Full-time	Full-time

higher level, the competing interests of the project and functional managers to ensure that the overall institutional benefits are realized and that individuals do not become mired in one role for too long. Having regular meetings between project and home organization to coordinate personnel assignments, as well as ensuring that project-assigned individuals participate in departmental staff meetings and technology review functions such as conferences, will help to maintain institutional balance and personnel integration.

21.10.4 Skunk Works

A so called 'skunk works', taking its name from legendary Lockheed advanced development projects such as the U-2 and SR-71 Blackbird aircraft, describes a group within an organization that is given a high degree of freedom, unfettered by standard bureaucratic processes, to generate breakthrough developments. The implicit connotation of the term is that of minimally constrained budgets, a freewheeling engineering team, and freedom from management. This section will discuss some of the issues encountered when managing such a project.

The principal advantage of skunk works programs is that a small, highly skilled and experienced team, left largely to its own devices and with minimal management oversight, can rapidly explore new technologies and develop prototypes. The operative phrases here are *highly skilled and experienced* together with *minimal management oversight*. Highly skilled and experienced implies that the team has the deep reservoir of technical talent relevant to the task at hand, and an innate understanding, after many years of developing technologies and products, of many of the issues previously discussed in this chapter. Minimal management oversight implies that the team is capable of exercising successful self-management, which requires not only understanding the

changing level of rigor as the development proceeds from low to high TRL, but also understanding and managing the complex team dynamic that inevitably results from the interplay of individual personalities. The entire team must understand and be committed to these self-imposed disciplines. It should also be noted that 'minimal management oversight' does not mean no management oversight, but rather that the management oversight process is streamlined and engineered to assist rapid development.

The principal disadvantage of skunk works programs is that skunk works teams rarely meet the fundamental criteria outlined above: i.e. being *highly skilled and experienced*, along with operating with *minimal management oversight*. Skunk works teams are often constituted imperfectly, with individuals lacking the necessary expertise and/or experience, and with management rarely creating optimal conditions for streamlined development. Consequently, more often than not, products from these efforts fall well below expectations.

The challenge for technology managers in skunk works environments is to select the appropriate team, set appropriate standards, keep upper management at bay, and successfully deliver breakthrough technology and product prototypes on a short schedule for affordable budgets. Inasmuch as the team is by definition isolated from the rest of the organization, and possibly from its industry as well, and perhaps in secret, it is incumbent upon the technology manager to maintain organizational interfaces and keep abreast of any technological or product breakthroughs in these other arenas. If the duration of the project is short, as is usually intended, this will not be an issue, but if the skunk works project is large and long term, other arrangements will need to be made. Beyond this, the technology manager needs to be acutely aware of team interpersonal relationships, technical progress, and the myriad daily issues arising that could derail the team. Management styles are different and a good manager will adapt the approach to the situation at hand, but there is no substitute for

being intimately familiar with the team's day-to-day activities and being willing and able to intervene as necessary to keep it functioning smoothly.

21.11 Technology Task Management

Managing a technology task requires the same fundamental elements as managing any other type of development task, i.e., defining the task objective, ensuring the development of clearly defined requirements, defining the technical approach, defining the task scope and success criteria, assembling and maintaining a competent and dedicated team, creating a work breakdown structure, developing and maintaining a milestone schedule, measuring progress, and ensuring the delivery of the final product within budget and schedule. This section, however, will focus on those aspects of technology task management that are different from other development tasks or are of special significance to technology tasks, and it will offer practical guidelines to successful technology task management based on the authors' experience.

21.11.1 Task Management/Leadership

Perhaps the primary task of the technology manager is to sell the technology and keep it sold. The practical truth is that technology development costs money and, without funding, the task will wither and die. While technology 'push' is often spoken of, the reality is that this motivator is usually not sufficient to sustain funding. Rather, funding of technology through the mid-TRL phase, and sometimes even the early-TRL phases, requires a 'pull' from stakeholders, coming from the capabilities that the technology promises to enable. From upper management's perspective, the task goal is not to develop technology for its own sake, but rather to provide a capability that has been defined as necessary or highly advantageous to the organization's business plan, strategic vision, or next generation product(s). Furthermore, investment in the capability(ies) under consideration must be made within available resources and provide an acceptable ROI.

There may be several technologies that are capable of providing the desired capability, and similarly there may be several different capabilities of interest to upper management. Thus, especially in the early phases, the technology task may be in competition with other technologies, other approaches to development of the technology of interest, or a completely different set of desired capabilities. The technology manager must be aware of this competition from both internal and external sources, and

periodically reassess the technology's ability to realistically provide the desired capabilities at the envisioned cost. Upper management will be monitoring the performance of the technology development team to determine if its development should be allowed to continue, and if so, at what funding level.

Most organizations cannot afford to maintain multiple competing technology development tasks throughout the entire development process, so it is typical to down-select at the breadboard or TRL 3–4 stages, where budgets are still relatively small. To survive beyond this phase and be promoted to the mid-TRL 'valley of death', the task manager must continue to persuade upper management of the capability or benefit that the technology is intended to provide, and to demonstrate adequate progress toward that goal. Therefore, the first order of business is to define the capability or benefits to be provided by the technology in specific and quantitative terms, i.e., provide metrics that can be measured and compared to existing or alternative technologies. Metrics may be in the areas of performance, reliability, cost, or ease of use, to name just a few; but whatever they are, they should be specific, measurable, amenable to demonstration in an incremental manner and, once they have been integrated, demonstrate the desired capability. Setting task goals and objectives accomplishes this.

21.11.2 Setting and Effectively Using Goals and Objectives

Goals and objectives are an often-misunderstood and contentious subject in technology development. A goal may be thought of as the desired end point, i.e., the desired capability or benefit to be provided by the technology that is being developed. It may be stated in either quantitative or qualitative terms, depending on who is defining the goal and the context in which it is being used. The technology manager must be flexible in this regard: to an immediate supervisor, the goal might be stated as a very specific quantitative metric, e.g., "*the technology being developed will be capable of providing 20 Gb/s bandwidth over 30 m for under 20 mW of power dissipation.*" While to upper management, marketing, or product managers, the goal might be stated as "*to enable a performance improvement that will leapfrog other available products by at least 30 % at no additional production cost.*" Objectives, on the other hand, are clear and unambiguous, and, when realized in their entirety, achieve the desired goal. They are sometimes described by the acronym SMART: Specific, Measurable, Attainable, Relevant, and Timely. In addition, objectives

should be capable of being broken down into sub-objectives that can be demonstrated in an incremental fashion so that significant progress can be shown as the task progresses through the TRL levels.

As discussed previously, it is critical to keep upper management enthusiastically supportive of the technology under development. To this end, the task manager must not only set objectives, but also stage demonstrations on a regular basis, showing the continuing advancement of the technology towards its ultimate goal by meeting the stated objectives. For example, one of the authors was involved in a technology task for which the goal was to revolutionize spacecraft avionics by developing electronics capable of reliably operating over an extreme temperature range, dramatically minimizing the need for thermal management along with its associated power, mass, and operational complexity. One objective was to develop an analog ASIC technology capable of meeting a set of required metrics for operational amplifiers and other analog components that could be designed using off-the-shelf tools, be produced by a commercial foundry, and exhibit at least a 10 year life while operating across a temperature range of -135 to $+185$ °C. The objective was broken down into specific sub-objective milestones, including the development of several foundational components, e.g., operational amplifier, switch, current source, and voltage reference, that met electrical operational requirements over the specified temperature range. As each component was designed, simulated, and then fabricated, it was tested and demonstrated across the temperature range and stressed for lifetime. With each demonstration, confidence in both the technology's ability to meet objectives and its ultimate goal was increased. Demonstrations were held frequently to keep the promise of the technology in front of upper management. The same strategy is useful in keeping the technology development team motivated and enthusiastic: frequent demonstrations motivate the team by showing that their hard work and dedication is paying off, that they are winning the technology development game, and that success is not only possible, but probable. It also serves to expose the team to upper management and the recognition that comes with management's acknowledgement of the team's achievements. Setting sub-objective-based milestones and using them to stage demonstrations that showcase accomplishments is an effective tool for the technology manager: in the above example, the series of demonstrations inspired not only the upper management but the technology development team as well. When problems came up, as they inevitably do when developing a new technology, the team was not deterred but remained motivated to overcome the obstacles and thereby show that they could succeed.

21.11.3 Defining the Technological Innovation or Advance

Goals and objectives are useful management tools, but technology tasks also require specific definition of the innovation or technology advance that is being developed—and that can often be quite difficult to define. For example, one of the authors was involved in developing a high performance computing system for space, based on the use of commercial off-the-shelf (COTS) components, and only after weeks of discussion did it become clear that the real technology advance was in the mid-level software that managed the computing system and in the analysis tools used to evaluate the radiation characteristics of the COTS components. All other elements, both hardware and software, were readily obtainable with no technology development. The focus of that technology development then became the development of complex COTS hardware component radiation test and analysis techniques and the middleware used to detect and manage radiation induced faults. Prior to that realization, the task was seen as being extremely complex with multiple interdependent facets. Afterwards, the task was sharply focused on two independent developments, and its tractability, efficiency, and cost were greatly improved. With the new focus, meaningful and sharply defined demonstrations could be planned and executed, and both management and the technology team could easily monitor progress.

21.11.4 Defining the Relevant Environment

In addition to defining the technological innovation, the technology task manager must define the environment in which the technology is to be utilized. The 'relevant environment', the key discriminator to progressing to mid TRL, is that environment which maximally stresses the technology advance in the first generation of envisioned products, thereby defining the bounds of the technology model that will be validated, the range of products that can safely be developed by using that model, and the conditions over which the technology is guaranteed to perform as predicted by the model. In the example above, while the technology advance was defined as a method of rapidly testing COTS hardware and a middleware layer to determine the error handling capability of the system, the relevant environment was defined as the computing environment; more specifically the computing system architecture, the application software, the real-time requirements of the spacecraft system, and the radiation environment that produced the maximum fault rate and the expected worst-case error set in the

first generation of computing systems that would use the technology being developed. In this case, thermal, mechanical, and electrical environments were not relevant, as these were well-known and well-understood factors that were not stressing to this technology advance. Similarly, compilers and software development tools were not relevant, as they constituted well-understood components that, while necessary, were not stressing to the technology. Clearly understanding and defining the relevant environment—and obtaining agreement of the same from both upper management and potential adopters—focuses the technology validation, facilitates its efficient management, and forestalls a myriad of questions about the technology adoption and use. While there are many factors that must be considered in deciding to adopt a technology, defining the relevant environment keeps the development team from having to address them all on a continual basis, thereby smoothing and streamlining both the development and the infusion process.

21.11.5 Other Technical and Programmatic Management Considerations

Obtaining funding and subsequently developing a technology through TRL 3 (the early technology development phase) is relatively straightforward; it is usually inexpensive and funding a task to this level does not constitute a significant risk to upper level management. Technologists can often execute this early phase with minimal attention to formal documentation, a rigorous technology model, the relevant environment, or verification and validation of the technology. At this stage the team need not be overly concerned with the eventual productization requirements of the technology adopters, or the rigors of technology infusion into real-world systems.

Transitioning to TRL 4–6 (the middle technology development phase) however, requires transforming the development team from ‘wild-eyed technologists’ to ‘hard-nosed product developers’. This transition is accompanied by higher levels of rigor; i.e. formal documentation, test methodology verification and validation, and attention to the envisioned product and system infusion path. Issues of manufacturability, testability, robustness, and ease of use become the focus in the latter half of this phase. It is at this stage that the task manager must help the team to cope with the internal pressures of comingling two very diverse cultures: those of the technology developer and the product engineer.

Internal frictions from the forced marriage of these two diverse—and in some ways opposed—cultures are inevitable and need to be carefully managed. At the intellectual level, education by acknowledging the respective strengths

and complementary natures of these two cultures can help to ease tensions. At the working level, socializing and encouraging common activities inside and outside the work environment can encourage the team members to see each other as individuals, rather than as ‘one of *those* people’. The importance of this type of activity should not be underestimated. As an example, some years ago one of the authors joined a small technology company where the fabrication process development and the manufacturing arms of the company were at odds—to the point of not speaking with—the design engineering arm. After observing the situation for several weeks, one day the author brought in bagels and set them down in a ‘neutral’ conference room adjacent to the CEO’s office, while asking the executive secretary to inform the technical staff that there were snacks in the conference room. This was done every Tuesday for several weeks. After a short time, conversation between the two sides became common, people spoke of non-work topics, and tensions were greatly eased. Tuesday bagels became a company tradition, continuing long after the author departed, and they continued to perform the function of allowing the staff to see each other as coworkers, rather than as hostile members of another tribe.

Setting milestones can also be used to help meld the team. Having the task manager define requirements such as the creation of detailed specifications, test procedures, test reports, and then explain what is required and why, helps the technologists to understand the need for the increased rigor and attention to detail contributed by the product development engineer. Meanwhile, as test, development, and debugging activities continue, the product development engineer will see the expertise and rapid analysis capability contributed by the technologist. The complementarity of their respective talents is thereby revealed and is able to be appreciated by the other side.

During this transition, upper management will also start to pay significantly more attention to establishing and achieving metrics and milestones, and pressure begins to mount on the team and on the task manager both from above and from within. The task manager, at this point, needs to buffer the team from upper management pressure while continuing to lead the team through the transition and refocus them on the new priorities. Keeping the team buffered from upper management pressure does not mean that the team is to be kept ignorant of upper management concerns, but rather that those concerns be communicated at the most propitious time; in fact, it is often beneficial to inform the team of these concerns and of the management’s increased interest in the development activities. This awareness can be utilized to motivate and instill in the team a feeling of both importance and urgency in the work to achieve the next milestone. It can also be used to promote team bonding, i.e., to create a common concern between the members of the two constituent team

cultures mentioned earlier. The manager's job is to utilize the increased attention from upper management, but to not allow it to create a highly stressing environment that saps the team's creativity, innovation, and energy.

Up to this point, the 'people' side of the technology manager's job has been discussed. It is now time to turn to the more technical, 'process' side of the management function. Because the technology development task is breaking new ground, the success of the task with respect to schedule and cost cannot be guaranteed. There is a need, however, to maintain a reasonable schedule and budget plan in order to keep the confidence of upper management and to ensure the continuity of the funding. At the outset it is also important to communicate the risks and uncertainties to upper management, as well as to instill confidence that these can be managed. Many of the strategies used for dealing with the technical issues of meeting quantitative scientific or engineering goals are the same as those for dealing with the programmatic issues of staying on budget and schedule. A technology task manager can profitably apply the following techniques.

1. Define a milestone schedule and development plan that increases the level of formality and imposes more rigor in the technology development as the team climbs the TRL ladder.

The formality and rigor required in documenting test procedures, test results, and analyses are much higher at TRL 6 than at TRL 4. Define what is needed at each TRL level, and let the team know what is expected of them as the technology matures. An attendant effect of the increased development rigor will be increased impacts to cost, schedule, and risk. Upper management needs to be reminded that maturing technology implies greater probability of eventual success but at greater cost; conversely, for the early TRL stage, where there is greater risk that an unforeseen problem will scuttle the effort, less mature technology implies lower probability of success but at lower cost.

2. Delineate the technical risks in the technology development approach and create a risk management plan with predetermined de-scoping options.

It is of paramount importance for the technology manager to identify and manage risk. For each risk, define (1) the impact of the risk, (2) a mitigation strategy, (3) a milestone at which the risk will be assessed, (4) the criteria by which the assessment will be made, and (5) the action to be taken if the risk is realized. Track these risks on a monthly basis, evaluating their probability and re-planning if necessary as the task progresses. Keep the team aware of the risk plan, looking out for unanticipated risks and encourage them to report new findings that could adversely impact the risk assessment and plan. Don't be afraid to update the plan and share the updates

with the team as well as with upper management. Furthermore, clearly define the budget and schedule impacts of the risk mitigation plan, as well as the technical ramifications of activating the 'fall back plan'. Making this plan early in the task takes the guesswork and the emotion out of the process, and is the sign of a mature technology task manager: it signals to upper management that the task is well managed and creates confidence in the team.

3. Have status and planning reviews on a regular basis to keep the management and the team informed.

In addition to the obvious fact that everyone likes to be informed, frequent status reviews help to ensure that issues and concerns are caught early—before they turn into serious problems—and do not surprise any stakeholders. Have status meetings at least monthly with upper management and at least weekly with the team; if part of the team is off-site, plan to visit them in person at least monthly, or if practicable, have the entire team meeting at their site periodically. As a matter of course, review the milestone schedule, the budget, the risk management plan, and recent progress against the task objectives.

21.11.6 Intellectual Property

It is critical for any technology organization that employs knowledge workers to identify and protect its intellectual property (IP), as well as to reward its employees for their innovations.

Traditionally, engineering notebooks, supplied and collected by the organization for archiving, are used to ensure that innovations developed over the course of the job are documented in a manner that is legally useful in future IP disputes. This practice should be mandated. In addition, it would behoove the organization to also have a computer-based system that allows technologists and engineers to quickly and easily note their innovations on-line, enabling management to catalog them and follow up as appropriate. Management should not only make their knowledge workers aware of these systems and encourage their use, but also work to persuade employees that documenting their innovations is of value to their careers and to the organization. In a similar vein, management should recognize and reward the inventor if the submission results in any type of IP protection action, e.g., patent application, copyright, identification as trade secret, or the like. Awards can be tangible, such as percentage of total royalty fees received by the organization from the invention, or intangible, such as a special designation of 'Chief', 'Principal' or 'Fellow' in the inventor's job title, or a combination of the two, but regardless should be distributed in a manner that is commensurate with the

value of the innovation to the organization. It is perhaps even more important that lesser value submissions be acknowledged and recognized: as mentioned previously, recognizing and rewarding inventors across the organization results in a focus on innovation, an awareness of the organization's need to innovate in order to compete and succeed in the market place, and a willingness to go the 'extra mile' to create new and exciting technology and products.

Acknowledgments The authors would like to express their gratitude to Carolee Kurta and Georg Siebes for their careful review of this manuscript. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

References

1. Mankins, J. C., "Technology Readiness Levels: A White Paper," NASA Advanced Concepts Office, Office of Space Access and Technology, 6 April 1995, <http://www.hq.nasa.gov/office/codeq/trl/trl.pdf> [retrieved 16 March 2012]
2. "FAA/NASA Human Factors for Evolving Environments: Human Factors Attributes and Technology Readiness Levels," DOT/FAA/AR-03/43, April 2003, <http://www.hf.faa.gov/docs/508/docs/TRL.doc> [retrieved 17 March 2012]
3. "Technology Readiness Assessment (TRA) Guidance," Assistant Secretary of Defense for Research and Engineering (ASD(R&E)), revised 13 May 2011, <http://www.acq.osd.mil/ddre/publications/docs/TRA2011.pdf> [retrieved 18 March 2012]
4. "Strategic Readiness Levels: The ESA Science Technology Development Route," ESA Advanced Studies and Payload Division, revised 13 Feb 2012, <http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=37710> [retrieved 16 March 2012]
5. Capon, N, "Managing Marketing in the 21st Century, Chapter 9: Managing the Product Line," p 317. Wessex Inc., Bronxville, NY, 2009.
6. Capon, N, "Managing Marketing in the 21st Century, Chapter 9: Determine and Recommend Which Markets to Address," p 190. Wessex Inc., Bronxville, NY, 2009.
7. Drucker, P.F. "The Discipline of Innovation," Boston, MA: Harvard Business School Press, Reprint R0208F, originally published May–June 1985.
8. Heilmeier, G. H., "Critical Questions for Research Proposals," Defense Advanced Research Projects Agency (DARPA), http://design.caltech.edu/erik/Misc/Heilmeier_Questions.html [retrieved 26 August 2012]
9. Christensen, C. M., "The Innovator's Dilemma: When New Technologies Cause Great Firms To Fail," Boston, MA: Harvard Business School Press, 1997.
10. Bolman, L.G. and Deal, D.E., "Reframing Organizations," 4th edn., Jossey-Bass, San Francisco, 2008.
11. Ancona, D., et. al., "Managing for the Future: Organizational Behavior and Processes," 3rd edn., South-Western, Mason, OH, 2009.
12. "A Guide to the Project Management Body of Knowledge (PMBOK Guide)," 4th edn., Project Management Institute (PMI), Newtown Square, PA, 2008.
13. "A Fast Technology Infusion for Aerospace Organizations," Shapiro, A.A., Schone, H., Brinza, D.E., Garrett H.B., and Feather M.S. Proceedings of the 2007 IEEE Aerospace Conference. Big Sky, Montana March 2007.

Robert J. Menrad and George W. Morrow

A project may have many forms—so many, in fact, that one may easily become confused when talking about them. So we must agree on what is meant when using the word ‘project’ and state how we use the term in this chapter so that it is clear how a particular process, technique, or consideration discussed here applies to your real-world activity.

The *Project Management Institute* defines a project as a temporary endeavor undertaken to create a unique product, service, or result [1]. Other definitions essentially convey the same meaning: an activity that results in an end-item deliverable, having a fixed beginning and a defined end. Although the National Aeronautics and Space Administration (NASA) employs the word project, other organizations in the USA use ‘program’ for the same activity. In this chapter, we will use the term ‘project’ and our agreed-upon definition, but we still need to discuss the specific kind of project to focus on. Although many projects are active at any given time, not all are the same. Some are straightforward while others are intricate and vast. Some require little funding while others require multiple years of high-level funding to sustain them. And some benefit from stable requirements while others undergo significant revision during their lifetime, perhaps even rendering them unrecognizable from the original vision. The best way to measure and describe these differences uses the concept of ‘complexity’. Simply put, some projects are more complex than others. For more than 50 years NASA’s Goddard Space Flight Center (GSFC) has specialized in project management for robotic scientific missions spanning the entire spectrum of complexity. With about 300 completed missions and one Nobel Prize for Physics [2], the project managers at GSFC recognize that similarities in definition

and activities only go so far and that, thanks to complexity, there is no short-cut or single recipe for success.

What is the project manager’s job? The authors have had the good fortune to know many excellent project managers, and each brings different strengths and skills to a project. One of the most experienced is Henry P. ‘Hank’ Wong. Hank has seen a lot of projects from the leader’s perspective while working in national defense, private sector, and civil space projects. Most applicable to the GSFC environment is his experience as a member of the first-generation NASA community involved in major programs, including Apollo and Skylab. This experience has resulted in the simplest definition for a project manager we’ve heard to date. As Hank puts it, the project manager’s job is ‘Getting work done through people’. Notice that it is not building a spacecraft, bringing a new technology online, or meeting budgets and schedules but *managing people in order to achieve these goals*. No single perspective addresses all facets of this position, but relationship-building clearly is a huge part of the role.

This definition sounds so simple that it appears almost too easy, but that is where the simplicity ends. Within any aerospace organization, teams of people are working on projects of varying complexity. Examples include component, box, subsystem, system, and ‘systems of systems’. Technology and space flight projects are sure to have many levels active at the same time. Yet, some will be successful, some will fail, and no two will be identical in their execution.

This chapter does not focus solely on a generic project’s internal characteristics because many project management texts cover this perspective. Nor does it focus on the internal characteristics of a technology project—the activity intended to develop a key new function or capability. Rather, it focuses on the relationship between a space flight project and a technology project. This critical relationship means to combine mission-validated capabilities with as yet unproven technologies in a way that results in a fully

R. J. Menrad (✉) · G. W. Morrow
Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Washington, DC, USA
e-mail: robert.j.menrad@nasa.gov

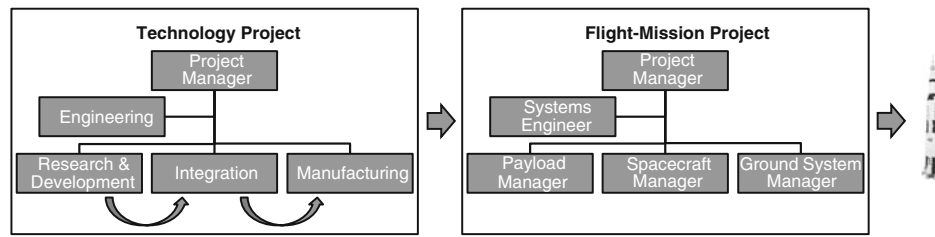


Fig. 22.1 Relationship of technology projects to space flight projects. The products generated by the technology teams are infused into the

space flight mission, which applies technologies to obtain measurements needed by policy makers and the science community

operational system ready for launch. GSFC's approach to projects for space flight missions will be the context for this examination. We know people in other environments may apply this relationship differently, but its underlying characteristics remain intact. Once a project manager and a technologist recognize these characteristics, their relationship matures and strengthens because of their increased awareness of one another's unique perspectives and needs. This awareness results in more accurate communication, which further supports the relationship, and the cycle continues to mature as it builds on itself. In this chapter we intend to define a recipe for a successful relationship between the project manager and the technologist—one that can span the spectrum of cultural norms, unique project attributes, and diversity associated with people who make up these two respective projects.

22.1 Project Types: Defining the Context

At any given time, many projects are underway within organizations that develop technologies—pure and applied research and development (R&D)—and apply them to obtain a suite of measurements from space-based assets. A technology-development activity is called the 'technology project'. Similarly, organizations involved in space-borne missions can have a significant number of 'space flight projects' active at the same time. Examples are

- Producing identical vehicles on an assembly line, such as constellations of identical satellites.
- Creating high-heritage vehicles with low-risk updates to subsystems, such as operational satellites.
- Developing high-risk, 'first-of-a-kind' applications of a detector, subsystem, or system, such as flagship science satellites.

In this chapter, we focus on the global perspective—emphasizing how the technology project's manager and the space flight project's manager interact. From our experience, this is the seminal relationship between these two

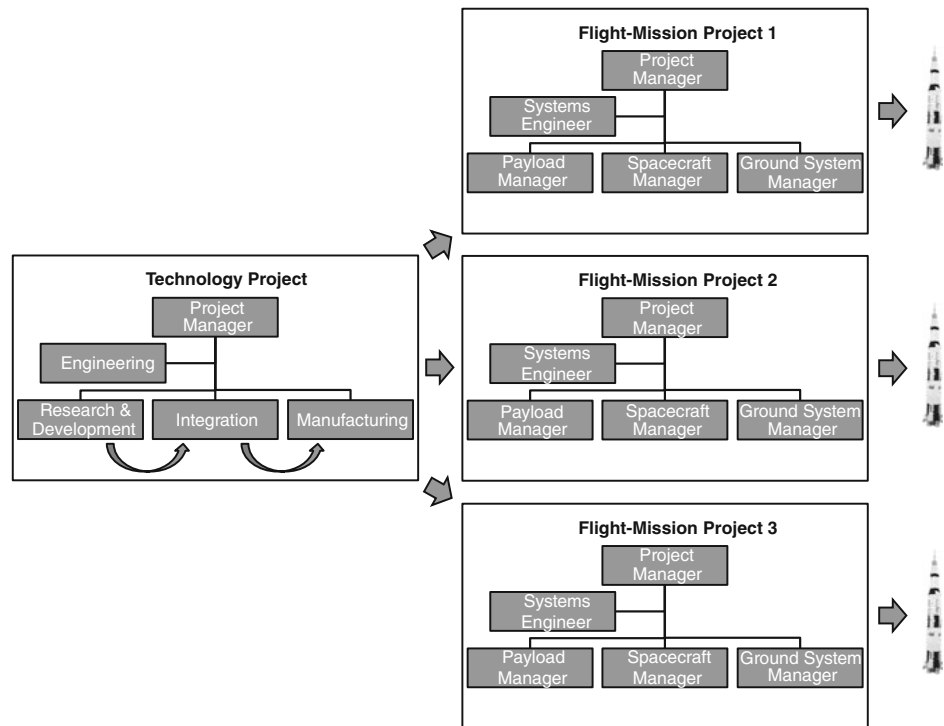
project types. It determines whether the original space flight project's baseline will incorporate the technology or keep it on the project when challenges occur. Figure 22.1 illustrates the fundamental relationship between technology and space flight projects. It describes each project's perspective, not how to do technology development.

A technology project has unique characteristics. In the simplest terms, it develops a 'product' intended to carry out a function or requirement. This can be a new piece part, such as a field programmable gate array (FPGA); a component, such as a detector; or a box or subsystem, such as ion propulsion. It can be a new approach for connecting various units to create more efficient interfaces, such as fiber optic data harnesses. The typical process is to conceive, invent, and improve the product over time, which normally results in updated versions or newer 'models' being released periodically. Each newer version usually represents some combination of corrections to errors in the previously released versions, new functions such as increased computing speed, added memory, or reduced power consumption, and increased reliability by applying better parts, more conservative de-rating strategies, or improved software and firmware.

We can measure the product's maturation using scales such as the technology readiness level (TRL) [3]; TRLs, along with key discriminators and exit criteria, are summarized in Table 21.1 and Fig. 2.10. For private-sector projects, we may protect the product using patents or other legal instruments in order to control its availability. When the product's value is high enough, we may apply national protection measures to prevent disclosing it or revealing its underlying design principles.

As the technology under development usually represents one aspect of a capability or function, the technology project has implicit limitations that the technologist must address. Most notably, the product being produced—no matter how impressive—typically cannot fly into space on its own. It must be integrated into a flight vehicle consisting of the payload suite and spacecraft bus.

Fig. 22.2 Perspective of the technology project manager. The technologist intends to develop a ‘product’ of universal interest to all potential customers



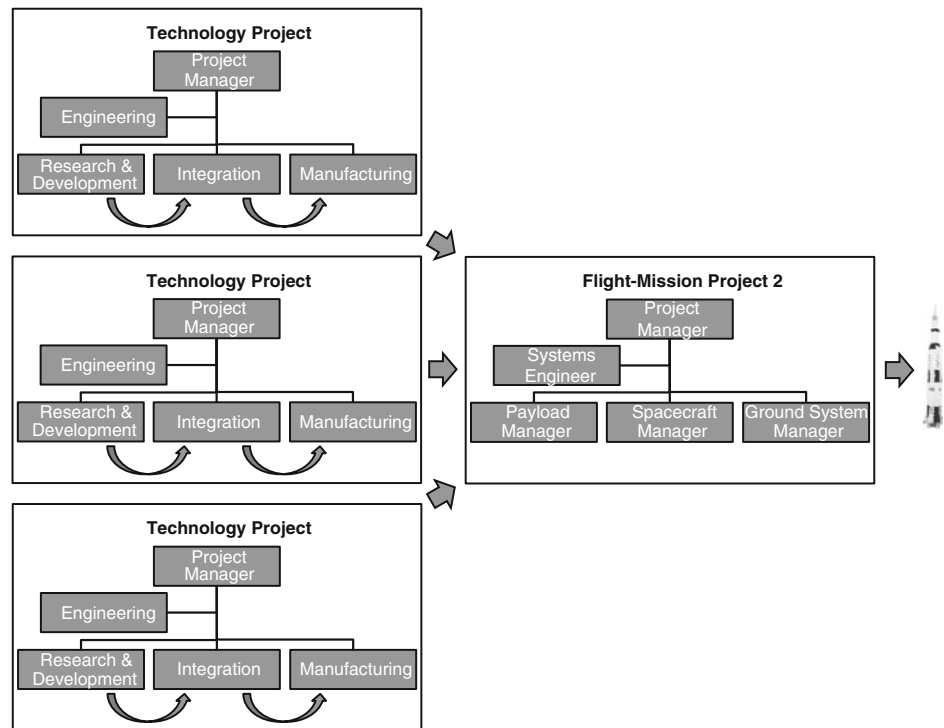
By contrast, the space flight project’s responsibility is typically the collection of data of interest to a specific community. This community may be policy makers, scientists, or defense agencies. In some cases, they will not directly care how this data is collected; in others, such as for the science community, they will be keenly interested. So the space flight project is interested in applying technology in order to assemble an end-to-end system. This system must be able to survive a space flight mission of known duration, under extreme environmental conditions, while taking accurate measurements for a predefined period (usually years). Whereas the technology project is interested in maturing a technology through periodic release of ever-better versions, the space flight project wishes to hold to a discrete engineering design that has been fully integrated and then expressed in terms of measurable cost, schedule, and risk. The technology project’s product is but one part of that discrete design. When a space flight project decides to apply a newer, immature technology, the development cycles needed to mature the technology must take place within a prescribed period, and that is a source of risk to cost, schedule, or requirements. Due to this risk, the space flight project manager is biased to seek out the most mature technologies available, except when the space flight mission intends to support the technology’s development itself.

As the technology generated must be integrated into the space flight project’s baseline plan, a technologist must either participate in a space flight mission whose express purpose is to fly the technology, or find a mission that will

incorporate it. In the first instance, technologists can drive the mission’s definition. In the second, they must ensure that the community of space flight mission developers is aware of the technology’s benefits at each point along its evolutionary pathway. They also must recognize that the technologist’s definition of ‘maturity’ is rarely the same as that of the project manager for space flight missions, and some part of the resources dedicated to the technology product’s development must be diverted to the space flight project’s needs. Project managers for space flight missions have the unique burden of ensuring that the technology community knows about the mission’s opportunities. The project manager must take advantage of technology forums that inform potential users on the products becoming available for use. This last aspect is key to the strategic communication approach employed by NASA’s Earth Science Technology Office (ESTO). Ultimately, we are examining the vastly different perspectives for technologists and space flight project managers, as shown in Figs. 22.2 and 22.3. The two managers must understand that both sides of this relationship are essential if they are to communicate effectively with each other and develop a successful collaboration.

Technology projects exist to satisfy requirements that are imposed by stakeholders or are specified when the technology is integrated into other projects. The latter might ‘improve on’ a product to fully exploit a technique, meet anticipated requirements of a future series of flight projects, or mature a proof of concept to allow developers to subsequently apply it with reasonable risk. In a few cases, the

Fig. 22.3 Perspective of the space flight project manager. The flight project intends to survey all potential sources of technology in order to design the most cost-effective system possible within resource constraints



product is fully mature and sitting idle, ready for application. Usually, though, maturation translates into successive versions of a product, with each new version having increased performance or functionality. By contrast, the flight project applies technology at a given state of development or with controlled further development. Technologists and project managers must avoid the natural desire to fly the most up to date version of the product possible. In short, everyone involved with the project must remember the mantra: ‘Better is the enemy of good enough!’

The rest of this chapter will shine a flight-mission spotlight on the relationship between the technology project manager and the space flight project manager while carrying out a space flight project. We intend to increase the respective communities’ awareness of the different perspectives, needs, and approaches in order that they can collaborate more effectively. For the perspective of the technology project manager and project team, see [Chap. 21](#).

22.2 Project Management: An Overview

As we begin to analyze the relationship between technologists and space flight project managers, we must establish common project-management principles. To do this, we will review their basics and then offer a new model for understanding how to carry out this critical competency. Everyone manages something, regardless of where he or she works on a project.

22.2.1 Understand Key Terms for Project Management

Many good textbooks are available on project management [4–6], but it is still useful to define basic principles.

- *A project*—Because the United States’ military space community refers to this limited activity as a program, we’ll apply NASA’s distinction between project and program in this chapter. A NASA project is an activity with a specific purpose and assigned start and stop dates. A NASA program is a collection of discrete projects with no explicit start and stop times.
- *Project team*—A collection of individuals, each representing a different competency, working together to complete the steps necessary to conceive, design, develop, test, and operate a space flight mission. Typically, this team includes project management, systems engineering, other engineering disciplines, safety or mission assurance, and project support.
- *Project manager*—The one person assigned overall responsibility for carrying out a project. This is a position, not the practice of project management. As an example, a generic flight project at GSFC can have at least six positions with the word ‘manager’ in their titles. As the project continues, each of these managers will practice project management but not to the same scope, complexity, or responsibility as that of the project manager. This is a crucial distinction. Project management occurs in many places, but only one person is responsible for the entire activity.

- *Project phase*—A collection of logically related project activities, usually culminating in a major deliverable or milestone.
- *Key decision point (KDP)*—The event at which the decision authority determines a project’s readiness to progress to the next phase or series of activities.
- *Formulation*—The phase in which a project
 - Develops a workable mission concept that would meet stakeholders’ technical, cost, and schedule requirements with acceptable risk
 - Assesses feasibility, technology, and concepts
 - Assesses risk, builds teams, and develops operations concepts and acquisition strategies
 - Establishes high-level requirements and success criteria, including life cycle cost and end-to-end schedule.
- *Implementation*—Carrying out the space flight mission’s stakeholder-approved baseline plan (the formulated project) as specified in requirements, budget, schedule, and risk terms.

22.2.2 Limit Overlapping Project Responsibilities

As the science of project management has become more formal and wide-spread, sometimes role inflation has taken place. This is most common between the project manager and the mission systems engineer—when the latter’s perceived scope extends into that normally reserved to the project manager. This confusion in roles can also occur between the technologist and the space flight project manager. Boards investigating mishaps have identified unclear roles and responsibilities as a contributing factor in accidents, so space-based missions must limit the effects of organizational complexities, such as role inflation. Once a flight mission intends to integrate a technology, individuals and teams must understand the clearly specified relationship between the technologist and the space flight project manager. Otherwise, strained relationships may challenge the collaboration, possibly even causing on-orbit failure.

22.2.3 Understand Project Management Frameworks

The Encarta Dictionary of English (North America) defines a framework as “*a set of ideas, principles, agreement, or rules that provides the basis or outline for something to be more fully developed at a later stage.*” Therefore, a framework is the construct by which work is done. Project management has several formal frameworks, as well as

informal ones based on the cultural norms by which the work is done. Formal frameworks are proliferating, but we will mention five common ones here.

National Aeronautics and Space Administration (NASA) Program/Project Framework—NASA’s Office of the Chief Engineer has created policy and guidance intended to standardize the activities for programs and projects across the five space centers and four research centers. This framework is specified in NASA Procedural Requirements, 7120.5e [7].

Project Management Institute (PMI®) Body of Knowledge (PMBOK®)—The PMI is based in the United States and has developed a project-management framework codified in the PMBOK and representing a generic 5-phase project. PMI is a non-profit organization dedicated to developing standards and guidelines. Its involvement spans research, continuing education, and publication of a magazine containing project-management topics. In addition, PMI has sanctioned strategies to ease networking through local charters, conferences, and training seminars. The generic project framework described in the PMBOK® is the foundation for a certification program. Practitioners earn the certificate by satisfying education requirements, gaining on-the-job experience, and passing a written exam. They maintain certification by meeting requirements for continuing education.

U.S. Department of Defense, Defense Acquisition Framework—The United States’ Department of Defense (DoD) has developed its own Defense Acquisition Framework (DAF) to complement the private sector’s PMBOK. It provides consolidated guidance and best practices for project managers who manage projects under the Defense Acquisition System. This system exists to manage the nation’s investments in technologies, programs, and product support needed to achieve the National Security Strategy and support the United States Armed Forces [8].

Projects IN Controlled Environments (PRINCE2®)—The PRINCE2 framework navigates project managers through all the essentials for running a successful project, as defined by the chartered organization. By design, PRINCE2 can be tailored to the needs of various projects. As with the PMI PMBOK, PRINCE2 is a generic ‘best practice’ tool. Government agencies in the United Kingdom and throughout Europe use it extensively as the recognized project-management standard.

International Project Management Association (IPMA)—The IPMA is a federation of project-management organizations based in Switzerland and chartered to promote international management while distributing common guidelines and standards across countries. This effort is important because the number of multi-national projects continues to

grow, as does the number of participant countries. The IPMA also seeks to establish professional standards and guidelines for project managers. This framework also contains a certification option.

22.2.3.1 Recognize What Frameworks Are and Are Not

A certificate is not a license to practice. As project managers lead increasingly complex projects, their proficiency must also increase. However, knowledge and certification, though laudable, are not the same as proficiency, which depends mostly on experience. Supervisors or decision authorities mainly use proficiency to assign project managers to projects that are more complex. After all the time and effort that is invested to become certified, however, it is easy to understand why managers might be confused about this issue.

Frameworks available worldwide are not projects, but they are very valuable to project management. They at least provide an excellent point of departure for a diverse project team who must begin by transitioning a generic process into a specific project. As long as practitioners are referencing the same framework, they can effectively communicate expectations using standardized terms, predefined best practices, documented attributes of completeness, and recommended tool sets. However, everyone must remember that no project will replicate a framework. Rather, managers tailor each project to meet the stakeholders' requirements, budget constraints, and schedule drivers.

Frameworks also are excellent yardsticks by which to measure a project's maturity. Organizations involved in projects over a long time amass a valuable historical record of what worked and what didn't. Project management maturity (PMM) is a process organizations normally use to evaluate a current project's maturity based on those past results. As presented in *Applied Project Management for Space Systems* [9], project managers should also be aware of a different approach for using PMM: regularly assessing the project (perhaps after key decision points) independently of any organizational activities. To apply this excellent practice, project managers should remember that frameworks contain widely agreed upon success criteria for their evaluations.

22.2.4 Responsibilities of the Space Flight Project Manager

Why do we need a project manager? Simply put, a project manager is the one person responsible for the group's efforts to meet stakeholders' requirements within cost and schedule constraints and with acceptable residual risk. This is true whether it is a technology project or a space flight project.

Although it is very easy to say this, the challenge lies in the execution. Table 22.1 illustrates many of the challenges that a project manager can expect to lead a project through.

22.2.5 Strategies in Project Management

Project managers must master what can be the most difficult part of project management: integrating different cultural approaches into a single team approach within differing environments. Otherwise, even a common academic understanding of roles and responsibilities or definitions will go for naught. Recognizing this challenge, organizations have defined alternate strategies for specific instances. These strategies in turn affect the project's environment. We examine three below: rapid prototyping, spiral development (introduced in Chap. 7), and project life cycle.

22.2.5.1 'Rapid Prototype' Approach

Technology development projects share a common desire: to take an idea and make it reality. The team's focus is on pure research and development. The very attributes that make good technologists—such as thinking 'outside the box'—often make them abhor structure.

A free-thinking environment is very empowering, especially when matched with a development approach that allows the team to focus more on developing the technology and less on paperwork, configuration control, risk management (beyond safety considerations), and so on. We are not saying all technology projects are unstructured, but the best ones often carry this bias. Because like-minded people tend to come together, this mindset is further reinforced in the technology community. Consequently, project management may become secondary to protecting the free-thinking approach.

Many free-form projects turn to rapid prototype development. In its purest application, this approach enables the project team to focus on developing the technology. Often termed define-build-test-repeat, it applies best when development speed is primary. Though empowering, this approach is far less structured and documented than other strategies, which means it may not match well when trying to link with those other strategies—especially the ones applied to developing human-rated systems.

Space flight missions typically use rapid prototyping differently: the project team assembles a system 'right now' to test concepts rapidly. This requires the project to employ current technologies because it does not have time to wait for a new technology to mature. In this case, the project will offer limited opportunity to test dramatically new technologies. Consequently, rapid prototyping normally applies to production programs.

Table 22.1 Project challenges. Any project experiences many diverse challenges

	Likely challenges	Brief explanation
Technical	Design deficiencies	Mistakes in design, analysis, or interface satisfaction discovered through analysis or test
	Component or subassembly generic failure modes	Failures requiring rework or redesign
	Poor performance	Failure to meet performance, availability, or reliability specifications
Requirements	Changes in user requirements	The 'moving target' effect: increased risk from incorrect mapping of changes to the baseline
	Changes to institutional abilities	Many projects leverage multi-mission abilities such as networks, WAN/LAN, and data archives. Unexpected changes to the functions they support burden the project
	Disconnects between planned versus actual functions of the technology devices	The technology plan is just a plan. When reaching a functional level takes longer than planned or becomes unachievable, the project must make up the difference. That will force new trades, designs, or implementation strategies. This typically happens when a project adopts immature technologies and hopes for the best. To solve this problem, we recommend critically assessing technology maturity levels
Budget	Incorrect phasing across fiscal years	Knowing when funds are required is just as important as knowing the funding level. Incorrect phasing of financial needs burdens the project and can constrain other critical activities
	Mid-year adjustments or changes (cuts)	External drivers are a frequent source of unplanned changes that usually delay the development schedule or increase risk. Categorizing the impact is critical
	Unplanned vendor adjustments	Because vendors are project team members, mission success depends on their success. When the business landscape changes, so do corporate expenses, such as rent and overhead. The project team must absorb these changes
	Technology maturation activities taking longer than planned	The time required to mature technologies is notoriously unpredictable. This means the project manager needs to plan for the most likely outcome, not the preferred one
Schedule	Constrained launch dates (generically, constrained 'finish' dates)	Many missions have constraints on launch, with planetary missions representing an extreme example. This reality requires detailed planning to maintain acceptable risk
	Workforce furloughs	Political considerations outside the project's control can strongly affect schedule
	Delays in technology maturation	A delay in technology development involves more than just increased funding requirements. Any delay can affect a project schedule's critical path. That's why technology shouldn't be maturing as the space flight mission develops unless stakeholders agree to this risk
Workforce	Skills mismatched to application	Space flight missions often represent 'one-off' applications of a unique technology. As a result the project must ensure access to specialized skills with enough people to make the progress assumed in the schedule
	Oversubscribed subject matter experts	Just as finding the right person can be difficult; getting adequate time from the experts is a risk we must factor into the schedule
	Attrition	The loss of any project member burdens on the remaining team members. Attrition can quickly affect productivity and quality, especially for the technology-development team
	Employee health	Employees commit themselves fully to many projects. As a result, stresses and pressures to deliver can take a toll on their health. The project manager has to manage the physical, mental, and emotional pressure as carefully as any other aspect

(continued)

Table 22.1 (continued)

	Likely challenges	Brief explanation
Logistics	Geographically distributed team members	Modern space flight projects involve participants from around the world. Managers must plan to bridge different cultures, time zones, and norms
	Natural disasters	Earthquakes, tsunamis, or hurricanes represent a fraction of the environmental factors that can harm a project. Although most projects can't afford to insure themselves, they must account for risk in order to limit their effects
	Non-disclosure constraints	Proprietary information constraints, national policies, or laws are examples of mechanisms that impede the free flow of technical information. Project plans must acknowledge these impediments

In some cases the challenge can be identified early and resolved, but in others the project, led by the project manager, must react quickly and effectively

22.2.5.2 'Spiral Development' Approach

In complex, multi-faceted activities with a lot of development, especially for software, information commonly evolves, so we need a different approach. For known work, the team applies focused, structured processes and procedures. At some predefined point in the future, they will begin work on defining the next wave of development, often leveraging the capabilities they have already started. In this model, the project is organized consistent with the structured needs of a space flight project but with integrated pockets of technology development. Rather than specifying discrete incremental releases or builds, the project defines development stages intended to complete part of the operations concept. Spiral development is very popular for 'systems of systems' spanning multiple phases, such as NASA's (now defunct) Constellation Program. It sought to build systems that could deliver humans to Mars by first demonstrating capabilities in low-Earth orbit, then at lunar distances, followed by the actual missions to Mars.

22.2.5.3 'Life Cycle' Approach

Huge 'systems of systems' programs are very exciting and can be a major user of new technologies, but they are far less common. Because rapid prototyping also is less frequent, stand-alone projects are more likely.

Organizations and their stakeholders have looked at collections of stand-alone projects as a 'portfolio' in order to address common strengths and weaknesses. By observing strong similarities across all projects in terms of how to conduct them, they have been able to break up the effort into a single set of discrete phases: the project life cycle. This results in significant advantages for the organization, stakeholders, and project teams. Table 22.2 summarizes the advantages of life cycle development.

Managers do not universally employ a single life cycle model, but each model shares a common key attribute: systematic development in which a project progresses through all of the steps that are required in order to be successful. Figure 7.6 presents the life cycle model that NASA uses now [10]; it's also the fundamental approach to project management at GSFC.

22.2.6 Putting Project Management into Practice

Books in the library and all development models on this subject emphasize one fact: project management follows a process. In general, the literature tries to propose an improved process, to explain an existing one better, or to compare and contrast the available frameworks. More important to our context, understanding the project management process enables technologists to communicate with greater clarity and collaborate more fully with project managers of space flight missions.

22.2.6.1 The Process of Project Management

The Heisenberg Uncertainty Principle teaches us that some pairs of data are mutually exclusive. In its formal presentation to a high-school physics class, the concept is illustrated using position and velocity. Typically, this is done during the examination of a picture of a bullet that has just been shot from a gun. In the lesson, we learn that the more we know about the bullet's exact position the less we know about its velocity. On the other hand, gaining velocity information for the bullet means we lose information on its exact position. We believe there is a corollary for mutually exclusive pairs of information in project management. For large-scope projects—those typical of space flight missions—knowledge of the overall project (horizontal scope) and specific detailed knowledge of any part (vertical depth) are mutually exclusive. Simply put, the project's scale exceeds any human's ability to know everything about everything.

This preamble prompts the question: What is a project manager to do? In the experience of the authors, the great project managers exhibit four key characteristics.

One: Being fully aware of horizontal scope. Project managers focus on knowing as much as possible about the entire project they are managing. Recognizing that they cannot know all the excruciating details, they focus on the people through whom they get the work done.

Two: Looking over the horizon. Anyone involved in a complex task naturally focuses on that task in a way that excludes all else; in effect, a form of tunnel vision. By

Table 22.2 Advantages of the life cycle model

	Advantage	Brief explanation
Project	Known expectations	Understanding the driving expectations on a project allows for more realistic planning, which in turn reduces risk and stress on the team
	Recipe for success	Less experienced projects now have a mission-validated method for completing the work
	Existing tools and procedures	Because work products for a specific life cycle phase are predefined, previous projects have developed the tools to support the them. This reduces the burden on the team to create new tools from scratch
Organization, such as hosting entity	Captured experience	Because the life cycle is based upon past efforts, organizations benefit from a proven method for doing the work. Thus, they can focus more on the unique drivers for this mission
	Scale for evaluating progress	With a single life cycle model each project is evaluated in terms of its maturity using the same criteria. This approach reduces ambiguity and allows organizations to identify issues sooner than would otherwise be possible
	Common 'language' across all projects	A single approach translates into a cultural norm for how to complete the work. This in turn allows sharing across projects, leveraging experience and perspectives to benefit all
	Common framework for evaluating overall performance, such as project management maturity	With a common framework, organizations can evaluate their overall performance across many projects and use the results to adjust personnel-development strategies, tools, and expectations
Stakeholders, such as an Agency	Commonality across all sources	With a common framework, stakeholders can leverage multiple sources more effectively and evaluate performance more accurately
	Universal statement of expectations in actionable terms	Higher-order activities, such as policy definition and garnering of political support, can now rely on a known set of work products having a given focus and level of detail
	Simplified insight and oversight	Oversight of an increased number of projects conducted by an increased number of vendors becomes more efficient
	Most accurate context for evaluating risk	With a common framework in effect, it's now easier to evaluate risks to program elements such as budget or schedule

A singular structured approach to projects results in benefits at every level

contrast, the great project managers are always examining the parts of their project in a systematic, round-robin fashion—taking a global perspective. As a result, they see things that require attention sooner than would otherwise be possible.

Three: Focusing on people as much as the work. Most project managers have technical backgrounds. Their extensive exposure to problem-solving in school and early career positions biases them to become task-oriented. However, at some point the position becomes less about personally completing a task and more about getting others to complete their tasks and produce certain results

- Addressing all necessary tasks.
- Scheduling so that completion of one task supports others that are in progress.
- Keeping costs for task completion within the planned budget.

Unfortunately, these skills form a much smaller part of formal training and early work experience. This is why leadership is such a challenging—and critical—element in a

project manager's career development. Again, it comes down to getting work done through people.

Four: Knowing the difference between good and great. Project management is not easy, especially for space flight missions. We have often told our teams that one cannot do this work without experiencing the proverbial 'fire'—but we just want a hose on every fire! So what is the difference between 'good' and 'great' when it comes to project managers in this kind of environment? The great ones know when to ask for advice or help!

Habits are hard to break, and project managers begin as discipline engineers, move to systems engineering, and eventually begin to manage projects. Therefore, project managers cannot assume they will have the same experience-based intuition in all of the project areas that they are expected to manage. This normally is not a problem until the trouble starts. When facing difficulties, we gravitate toward areas we know the most about. Further, we believe we can contribute best by identifying the root cause of a problem and then defining the best solution to address it. At

the same time, we may make the huge mistake of turning away from problems in areas we know less about. Project managers must evenly distribute their time across all project areas in order to keep a finger on the project's pulse; and, sometimes, to keep their jobs.

22.2.6.2 Why Projects Fail

Each time we embark on a project, we accept the risk of failure. By nature, a project manager is an optimist—willing to expend the effort required to overcome the risks in bringing the mission concept to fruition. However, despite the best efforts of those involved, projects do fail. Organizations, stakeholders, and academics have written many studies attempting to explain why this occurs.

All projects have a built-in bias for failure. Although many factors are at play, our experience shows that the key driver is the very optimism that prompts people to undertake these challenges in the first place. Optimism abounds on how long it will take to do the work, the stability of the budget or project team, and the requirements of the project. In fact, we are biased toward optimism in every underlying assumption that we are required to make when moving through life cycle phases.

Yet, complexity is as much a driver as optimism, if not more so, for why projects fail. This should not come as a surprise. Intelligent people typical of our community have developed grand visions for what is possible because of readily available computing power and today's sophisticated computer-based engineering applications. They may pay too much attention to maturing the project's individual technologies but too little to maturing the final flight system across all life cycle phases. Take NASA's original design for the International Space Station. It employed a double keel that was as grand in its vision during the conceptual phase as it was possible to 'engineer' using the computer models available at the time. Later, it became questionable whether the number of astronauts needed to maintain this version in orbit could be accommodated aboard the facility.

We use this example because it is straightforward and easily recognizable as an instance of complexity exceeding capability. Alas, in too many cases, the consequences of complexity are not so readily apparent. Table 22.3 presents some sources of complexity that project managers must factor into their plans. We want to provide a richer understanding of what project planning must take into account—the very phase when a technologist wants to influence the project the most. In addition, project managers often inherit an already formulated project, which means they had little to say about the way a project was set up. This list of complexity sources can help you, as a newly assigned project manager, in your first assessment at leading a project.

22.2.6.3 The 'BEER Model': A New Perspective for Success

Project managers need tools that make it easier to practice their craft. Many textbooks spend a lot of pages on what project management 'is' and others spend pages on methods. However, the 'BEER Model' captures the essence of project management, as we have observed it over many instances.

One antidote for project complexity is 'completeness'. By paying proper attention to all aspects of a project in both the scope and time domains, project managers increase the likelihood that their project teams are similarly paying appropriate attention. This action alone will not eliminate the risks, but it can greatly reduce them.

Then what is the project manager to focus on? What represents 'complete' attention to the project? How can we organize these activities so they make intuitive sense? The textbook *Applied Project Management for Space Systems* [11] addressed some of these questions. The unique group of practitioners and educators who edited this text recognized that they could aggregate most project manager core competencies into four discrete actions aligned with behavior. Specifically, project managers need to 'Bound' the project in ways that ensure completeness but don't exceed the stakeholders' needs, as codified in the requirements. They need to address the processes, procedures, and techniques needed to 'Execute' all required activities, as well as 'Enable' project team members to leverage their experience in ways that will make the whole team successful. Finally, and equally important, project managers must be willing and disciplined enough to 'Reassess' how the project team is performing—not on a specific task or series of tasks, but as that performance affects the project globally. When project managers think this way, they can see the project not just as a participant but also as its leader. Table 22.4 illustrates the BEER Model.

22.3 Project Management: A Collaborative Approach to Project Execution

Space flight projects are often large, visible, complex, and prone to failure. From the Goddard Space Flight Center's perspective, they're almost always one-time builds conducted under schedule and budget pressures, with probing stakeholders and managers often asking for status and offering solutions. To overcome these external and internal challenges, a project team becomes close and, if the project manager is successful, takes on a common persona that unifies them. Likewise, technology project teams are groups of people bound together by the common vision of

Table 22.3 Sources of project complexity

	Types of complexities	Potential impact
Technical	Scope (such as system of systems)	Scope drives the infrastructure of the process in terms of requirements traceability, configuration management, risk assessment, and so on
	Technology application	Using technologies not originally envisioned with the application can cause complexity that can span hardware and software
	Commercial off the shelf (COTS) applications	While offering an immediate savings, COTS issues include licensing, IT security implications, integration and obsolescence
	System segmentation	Systems such as ground data systems can have many elements. If we are not clever in grouping these elements, we may have to manage too many interfaces
	Software (flight and ground)	A science in its own right, managing the architecture, modules, and libraries increases the challenges of integration and maintenance
	Hardware (flight and ground)	Challenges in detector electronics are classic examples of anticipating complexity but yet underestimating its effects
Organizational	Cultural norms (national and otherwise) of the organization in which part of the project is located	Established formal and informal practices can drive projects to carry out more activities than would be normal given the level of risk assumed to be acceptable
	Stakeholders	The sheer number of stakeholders to be briefed can challenge a project's ability to execute within schedule and budget assumptions
	Vendor proliferation	Having many contractors, subcontractors, and suppliers burdens resource managers
	Human interfaces	Depending on the nature of the project, the user community alone may be more diverse, resulting in more representatives to meet with regularly. This is especially true when missions involve atypical components, such as nuclear power subsystems
	Responsibilities, authorities, and accountabilities	All teams require a clear definition of each member's role and responsibilities. For large projects, we may need a lot of time to define them and then to modify them based on performance
	Proprietary data	Protecting national and corporate proprietary data can require more from management information systems and special training for the project's core members
Environmental	Weather	Floods, hurricanes, or tornadoes are just some examples of events that can have a long-term impact on a project. In this context we're focusing less on the risk of occurrence but more on the consequences of resolving them, such as backup control centers
	Natural disasters	Institutional facilities—such as ground stations, network hubs, and points-of-presence—carry data of varying priority, which means we need diversity in the implementation
	Political climate	Changing priorities at the national level can cause basic changes in the support for or against a project.
	Working conditions	Obsolete buildings, unsafe conditions (such as asbestos), or delays in acquiring new or modified facilities

developing something new and then seeing it flown successfully. They also develop their own persona. How they work together strongly influences any collaborative effort's ultimate success. The key to success is in how each team sees the other. The technology team should not consider the space flight project as a means to an end, nor should the space flight project consider the technology team a miracle cure for a poorly defined project. To integrate these teams successfully, each manager must understand the other's project (Table 22.5).

Outside GSFC, some projects focus on production builds or other applications of similar complexity. Although this production setting and GSFC's one-of-a-kind environment have inherent differences, they also have similarities, such as inserting technology updates or new technologies into development. An excellent example is the satellite phone industry, which excels in mass-producing flight systems that evolve with new technologies. In any case, the relationships linking technologists and project managers must work well in any environment.

Table 22.4 The 'BEER Model'

	Focus or goal	Model aspects	Brief explanation
Bound	Define the project's foundation in terms of user needs	Project capture	In a competitive environment new business capture is itself a project with special considerations we must account for
		Project planning	The project plan is the focal point for project governance; it establishes procedures the entire team must follow
		Organizational design	How we organize our team directly affects the way members are empowered to carry out their assigned responsibilities
		Stakeholder management	Every project must correctly identify the stakeholders and how they affect the project, so we can accurately apply ways of meeting their needs
		Management information system	We must carefully consider the methods and systems used to control the project's collective knowledge, especially as complexity increases
Execute	Assemble a project's building blocks to create an environment for success	Systems engineering	The practice of bringing various disciplines together into a single, integrated system that we can specify in terms of its construction, margins, and performance
		Requirements traceability	The process of defining a system in functional terms across multiple levels of detail
		Logistics	Approaches to managing the operational systems
		Mission assurance	An engineering-based discipline focusing on issues of system reliability, safety, and satisfying requirements
		Life cycle cost estimation	The process of estimating the funding required for a specific system of given complexity from start to finish
		Budgeting	The process of managing a project's funds with focus on multiple years of performance, cost plans versus new obligation authority, reserves, and resource-loaded risk analysis
		Acquisition and contracts	Procurement considerations, best practices, and constraints
		Monitoring, evaluation, and control	The practice of observing project performance in terms of people and products, all with the intention of identifying issues as early as possible so we can retire them before they affect the project baseline
		Risk management	The process of identifying issues before they occur so we can mitigate the event or its impact
		Software development	Dealing with the unique challenges of developing and maintaining source code in firmware, microprocessors, and systems
		Hardware development	Dealing with the unique challenges of developing physical systems that meet requirements for availability, reliability, and time to repair
Enable	Focus on the project's human aspects	Leadership and teamwork	Project managers apply leadership: a skill that requires nurturing as much as any other
		Strategic communication	Building advocacy for our projects is critical to survive periods of fiscal austerity or technical challenges
		Critical decision making	Everyone on a project team makes decisions, but because the project manager makes the most critical ones, this skill must be developed
		Culture	Leveraging the cultural norms every organization has is easy when they help the project, but how do we deal with those norms when they create hurdles?

(continued)

Table 22.4 (continued)

	Focus or goal	Model aspects	Brief explanation
Reassess	Step back to periodically re-evaluate the project's progress and overall efficiency	Project management maturity	By evaluating our project against a predefined scale, we can identify where the team is behind in the 'maturity' of its work and apply more resources to address the deficit
		Strategic issues in project management	Continuing education is a critical part of the project manager's work portfolio. Sharing knowledge and lessons learned with other practitioners ensures we are not blind to the strengths and weaknesses of an individual or the team. It also ensures we apply corrective measures

A model describes aspects of a given discipline. The BEER model describes the practice of project management in terms that enable its practitioners to address this complex activity more completely

To collaborate well, a technologist must understand the project manager's perspective. In technology projects, he or she often is the project leader and the one who had the idea in the first place. Depending on the situation, the technologist may have grown that project into laboratory status, with multiple missions under the team's belt. Alternatively, they may be just starting out, looking for the first great opportunity to apply the technology. In both cases the technologist is almost always smart, a gifted engineer or scientist, and a visionary. Often, he or she is an excellent marketer interested in having the technology applied in as many ways as possible.

Space flight project managers share many of these attributes and we have found them to be equally smart, gifted engineers and visionaries. However, they are decidedly not marketers. In fact, they listen to claims about the next 'big thing' with a lot of skepticism because they need to maintain their project on the lowest-risk path to its conclusion.

Technologists must recognize this mindset as a stable characteristic of project managers, knowing their skepticism will rise whenever claims increase, the project gets closer to its delivery, or its budget becomes more constrained. This means technologists must understand that the space flight mission is not an opportunity for 'pure' research—it is 'applied' research and development. Only in this way will the two professionals collaborate harmoniously.

22.3.1 Spanning the Project Life Cycle: Defining Common Threads

Project management is iterative, so project managers and technologists spend their management time iterating through many activities that take place in parallel. This method is similar to a computer operating system's round-robin approach to servicing the multiple processes running on it. Some of these activities are specific to a life cycle phase and, once completed, are seldom revisited or are simply updated. However, we do carry out many activities throughout the project regardless of its life cycle phase. They are the 'common threads' of project management.

As we turn our attention to collaborating on a project, we will focus on these common threads between the technology and space flight mission projects while emphasizing the different perspectives of technologists and space flight project managers. Armed with this knowledge, technologists are better positioned to tailor their messages and products so that the space flight project team can more readily accept them. This in turn ensures the most favorable consideration of the technology when the team develops a mission concept, integrates the technology during the implementation and integration phases, and includes it in successful mission operations.

22.3.2 Recognizing Organizational Knowledge and Culture

Organizations want to reduce uncertainty about their processes in order to extend their own longevity. This means they will repeat what has been successful. They capitalize on their successes by using written and unwritten rules to inculcate their processes into the workforce's collective knowledge. These rules become the basis for their culture and are commonly referred to as 'cultural norms'.

Cultural norms drive environmental factors that greatly affect space flight projects and, by extension, the relationship between project managers and technologists. Therefore, technologists and technology projects should take the time to understand a space flight project's organization and cultural norms as much as possible. This understanding will be much easier if the technology project is part of the same organization. However, a lack of understanding can harm relationships across national borders, business sectors, such as government and civilian space, and stand-alone elements in a large organization, such as between NASA's field centers.

22.3.2.1 Understand that Cultural Norms have Implications for Project Workforces

Cultural norms continually constrain project managers throughout a project's life cycle, as long as the environment doesn't change through reorganization, merger, or departure

Table 22.5 Project roles and responsibilities

Project position	Responsibilities
Project Manager	<ul style="list-style-type: none"> • Leader of the project team • Ultimately responsible for project team's performance and of the mission's operational performance • Ultimate authority for budget and commitments • Is 'owner' of the project plan: the one person responsible for this document's content, as sanctioned by the governing authority
Deputy Project Manager	<ul style="list-style-type: none"> • Leads the project team when the Project Manager isn't available • May be the contracting officer's technical representative for U.S. government contracts, which results in significant responsibility. But the Project Manager specifies actual authority • Often leads special teams in focused efforts to support the project or address a problem
Deputy Project Manager/ Resources	<ul style="list-style-type: none"> • Leads the resources management team • Ultimately responsible for the integrity of the budget process and expenditure of funds • Key negotiator for the project; requires the most detailed information on direct and indirect cost billable to the project • Leads the project and its partners and contractors in an annual budget process • Defines the process for requesting access to project reserves, though the Project Manager makes all final decisions to encumber the reserves • Is 'owner' of the basis of estimate, the integrated baseline, and all performance measures, such as earned-value metrics
Mission Systems Engineer	<ul style="list-style-type: none"> • Leads the full engineering team • Is the final authority for determining what trade studies are required, when studies are completed, and how to resolve engineering issues within the team • Ultimately responsible for the mission's technical performance • In some organizations holds technical authority bias over the project independent of the Project Manager • Is the 'owner' of the system specification and systems engineering management plan
Instrument Systems Manager	<ul style="list-style-type: none"> • Manages the overall payload activities • Manages the instrument managers • Is the 'owner' of the payload-level interface specifications
Observatory Manager	<ul style="list-style-type: none"> • Manages how the team defines, builds, and integrates the spacecraft (the part without the instruments) • Is the contracting officer's technical representative for the spacecraft manufacturing contract
Ground Systems Manager	<ul style="list-style-type: none"> • Manages how the project defines, implements, and transitions the ground segment to operations • Responsible for designing and building the project's unique ground-based computing systems, such as ground network apertures, flight simulators, and data processing and archive centers • Is the 'owner' for the ground system requirements, implementation plan, interface control areas (ICAs)/ interface control documents (ICDs), flight databases, and procedures

Each project has senior-level positions it must fill with practitioners who can carry out their assigned responsibilities through the efforts of assigned team members

of key employees. These constraints are extensive—spanning workforce groupings, individual practitioner roles and responsibilities, internal and external stakeholder expectations, and required technical, schedule, and cost margins. They may also define the accepted suite of applications and tools by which projects do their work, report, and publish products. One of the most significant implications is how much direct control the project manager will have over the space flight project and the people assigned to the project team. For example, suppose an organization groups its staff by product line, capability, or function. In this case, a dedicated manager will supervise people and the

development effort, so the project manager may have little control (other than persuasion) over team members. He or she typically becomes a key data source by acting as a bridge between the end-user and the development group.

At the other end of the spectrum, some organizations 'project-ize' their workforce by grouping people into discrete projects, with a project manager controlling the project, team-member assignments, and their performance assessments. Although these organizations give project managers maximum control, they exhibit inherent weaknesses, e.g. requiring time to hire new project members, the planning required to develop employees, completing

performance assessments, with corrective action in some cases, and placing personnel on new teams when the project is completed.

To be successful, project managers need skills in managing human capital and in the disciplines that a project requires.

Most organizations have sought a hybrid solution that leverages the strengths of these two approaches while minimizing the inherent weaknesses: i.e. a ‘matrix’ organization. Here, employees may be organized by end product, function, or capability but are assigned to the project manager while on the project. Matrix organizations use this concept to varying degrees and give the project manager varying amounts of control.

Technologists must understand the project manager’s control level in order to adjust the marketing of their technology and integrate the two projects more smoothly once they have decided to collaborate. Refer to the Project Management Institute’s Body of Knowledge [12] for details on organizational structures and their implications for the project managers.

22.3.2.2 Follow Two Key Principles for Managing Cultural Norms

First, know the environment. As a technology manager, if you’re not in the same organization as the space flight project, you must bridge its cultural norms. Even in the case where both projects are in your own organization, other departments may have different cultural norms. This is especially true if the organization has acquired other companies and is merging different cultures. You must spend time identifying and then evaluating the differences between your own environment and the space flight project’s culture that will drive development. If you do not bridge cultural norms, you may discover mismatched expectations, policies, and processes that turn what could have been excellent collaboration into a stressful relationship.

Second, do not expect latitude where none exists. For example, you may have spent time in the freedom of an academic or loosely governed organization and now want to see the cultural norms of your home organization applied to your part of the space flight project. Wouldn’t it be nice if it worked this way? It doesn’t. You have to anticipate how much flexibility the project manager’s home organization will allow you. This consideration is important because these norms will affect your entire technology-project team, control what work is needed, and determine how you will report status. Therefore, you must anticipate how team dynamics may affect your people. In addition, if you are following a legal contract that establishes your relationship with the space flight project, you will need to be sure it reflects the potential inefficiencies of this new environment.

22.3.2.3 Understanding Project Roles and Responsibilities

Many people have studied teams, their characteristics, and the tools needed to measure their effectiveness. Ideally, leaders want their teams to be interdependent and cohesive. Interdependence means the members’ skills have minimum overlap, everyone can communicate openly and candidly without retribution, and each person can rely on the others to complete tasks in time for their own work to progress. Cohesiveness measures the strength of ties that bind individuals to the team, keep them motivated, and commit them fully to continue working towards the project’s assigned goal. As the technologist, you must apply your own measures to ensure that your project reaches the highest level of interdependence and cohesiveness. The project manager you are collaborating with will do the same.

Unless a contract determines the relationship between the space flight project and the technology project, collaboration will depend on how effectively the two teams work together to achieve their fullest potential. For two groups to collaborate effectively, though, each must understand the other’s roles and responsibilities. Therefore, you must know the roles and responsibilities of the core members of the space flight project if the right discussions are going to take place. Not recognizing constraints on a person’s position for making agreements or committing resources can set both parties on the wrong path, which results in lost time, wasted resources, and lost trust. Figure 22.4 illustrates the structure of a generic space flight project in terms of its segments.

As Fig. 22.4 illustrates, the project’s positions relate to each other hierarchically, so each position has a span of control or authority—normally referred to as roles and responsibilities. Figure 22.5 defines the roles and responsibilities of a project’s senior members on a typical space flight project.

22.3.3 Managing Stakeholders and Their Expectations

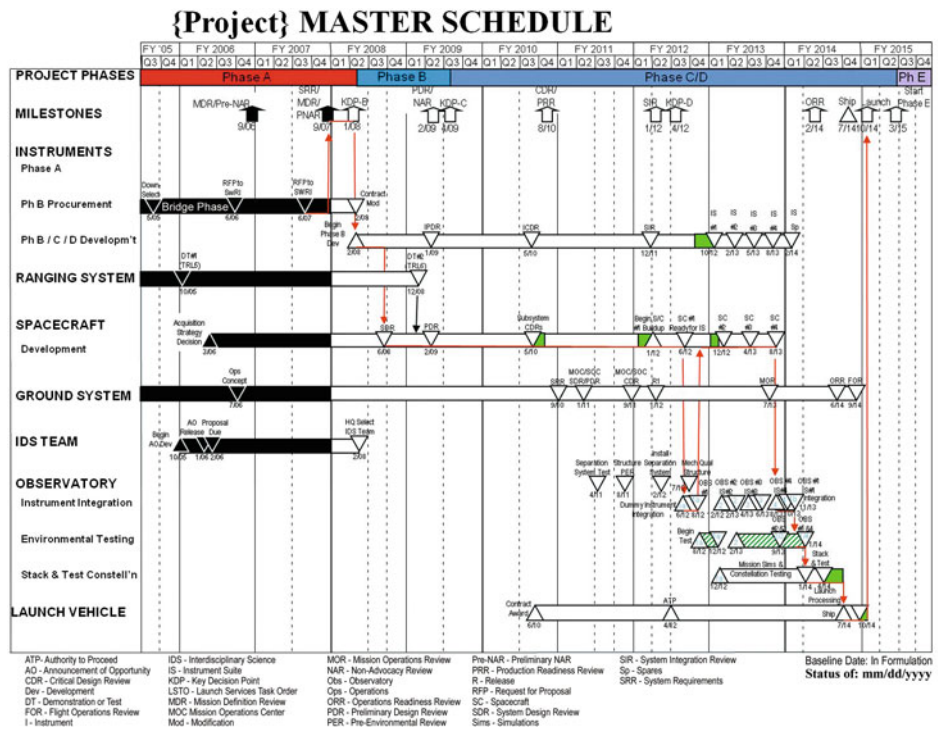
To this point, we have discussed the environment created by the organization that is hosting the space flight project and then focused on the project’s generic structure. Our third common thread continues the focus on people, but expands our perspective to the mission stakeholders.

By definition, stakeholders are individuals or organizations that can directly or indirectly influence a project. They can advocate or oppose it and be passive or active in their involvement. That means they can affect what a project does, how it does it, and what ultimately defines mission success. On a space flight project stakeholders can approve

Fig. 22.4 Generic space flight project. A space flight project is a hierarchy made up of people spanning multiple disciplines—all working together to achieve a single goal



Fig. 22.5 Top-level project schedule. A typical integrated mission schedule (IMS) has thousands of schedule items in its database. The top-level schedule boils it down to a scale that makes main trends clearly visible



waivers to excuse process requirements in order to achieve simpler execution, or they can extract every last ounce of effort from the team in terms of documentation, analysis, and requirements verification. Simply put, their requirements can influence a project’s workload, budget, and schedule for good or bad.

Stakeholders often overlap for technology and space flight projects but not always completely. Therefore, a key core competency for project managers and technologists is

the ability to identify stakeholders, categorize them by their ability to affect the effort, engage them in a timely manner, communicate with them effectively through direct interaction or other methods, and ultimately manage their expectations while carrying out the mission.

Whenever project managers and technologists work effectively together to construct a strategic communication approach for key stakeholders and use it regularly, fantastic results can follow.

22.3.3.1 Identify Stakeholders

A lot of academic work has considered stakeholder management, including how to identify them, categorize their influence, and communicate with them. The PMI PMBOK [13] is an excellent example of a project-management framework that includes this topic. However, one can easily be consumed by the thesis-like discussions of ‘layers’ of stakeholders, the names of these layers, the nature of the stakeholder’s influence on the project, and the process to interact with them. For success in this area, we need to examine the topic broadly enough to ensure that we have a complete list. At Goddard Space Flight Center, common stakeholders include the Center Director and senior staff, Directorate heads, the Principal Investigator or Project Scientist, and line managers with matrix personnel who join the project as and when necessary. Notice that we have not even left the center yet! Add in managers at NASA Headquarters, corporate managers of the project’s partners, and so forth, and we see that the list grows very quickly.

No matter how much effort we put into identifying stakeholders, we are likely to miss some. Which ones? Probably those with a more tangential relationship to the project, e.g. people supporting the U.S. National Environmental Protection Act (environmental impact), Department of Energy (nuclear power), Congressional staff (programmatic), other governments (international missions), environmental groups (from nuclear power systems to the shipping of hazardous waste), and local community action groups.

To help resolve the risk of missing some stakeholders in the original survey, project managers, technologists, and their respective teams regularly reassess their lists. As flight projects change from their baseline, we are likely to remove some stakeholders, add new ones, and occasionally change how we engage those that continue.

22.3.3.2 Manage Stakeholder Expectations

Whether representing a technology project or a space flight project, technologists and project managers talk to stakeholders from a position of integrity. This requires each project team to deliver current, clear information on the project’s status and whether it will continue to progress and meet commitments in the near-term. When a space flight project adopts one or more technologies, the project manager will include each technology’s development status in reports to the stakeholders. Technologists must recognize the importance of stakeholder management and fully engage in it.

22.3.3.3 Avoid Stakeholder Conflicts

In dealing with some of their stakeholders, a pattern became obvious to senior managers at GSFC and the Jet Propulsion Laboratory, namely that mission success was not as easily defined as they had assumed. Sometimes, stakeholders did

not agree with management concerning the observatory’s ‘successful’ on-orbit performance or operational status of a ground-data system element. Both centers sponsored a study in the early 1990s to understand how success can be considered failure and vice versa.

The detailed results of this study appear in [Chap. 22.2 of *Applied Project Management for Space Systems* \[14\]](#), so we will not repeat them here. However, we do want to emphasize that we cannot evaluate all projects using the same success criteria. Rather, the study reported a new concept—mission paradigm—that applies to all projects. This means projects may have science, technology-demonstration, or operational missions, and the stakeholders’ expectations and criteria that define success will differ for each one. So technologists and space flight project managers cannot assume their definition of success—build it no matter how long it takes and have it work perfectly in orbit—is the same as the stakeholders’ definition. Therefore, project managers must learn the stakeholders’ expectations when the project begins in order to clearly define its mission paradigm. Only then will project managers and teams know how to accurately weight the engineering options, assess risks, and meet expectations.

Technologists and the project managers must have an equally candid conversation when discussing whether to adopt a new technology, because the paradigm will also extend to the technologists’ activities. For example, on technology demonstration missions, a technologist may be able to influence stakeholders into extending schedules or incurring more cost. However, on science missions which use that technology to collect data, the same influence may not exist. In fact, this example highlights one of the most important distinctions between the teams collaborating on a mission. While one team focuses on meeting schedule and cost, another may work to meet functionality. Clearly, this disparity is not in the best interest of the mission or the technology, so teams must agree on a common focus before the challenges begin. Otherwise, manageable problems may become unmanageable.

22.3.3.4 Handling Standards and Guidelines

Experienced stakeholders who have acquired a lot of historical performance data and worked on many projects at the same time commonly write down their expectations in advance. Known as policy documents, standards, or guidelines, these documents specify the expectations and constraints for the space flight project throughout its life cycle.

These expectations and constraints can be extensive, so the documentation will span all aspects of the project. Examples are applying engineering practices, such as tracking technical performance factors, assuring quality, such as frequency of inspections and audits, keeping people safe, protecting national assets, establishing budget

reserves, documenting confidence levels for cost estimation, and handling mission risks.

At certain times, we may modify any of these documents in order to address a need, such as showing insights gained by the organization, changes to risk tolerances, or adjustments to affect overall performance by projects at the portfolio level. Relief normally comes to the project only through an established waiver process—with no guarantee that a given criterion will be eligible for waiver.

Technologists must be prepared to work within these established norms, and most are ready to do so. In some cases, we must adjust the technologist's deliverable, which in turn requires retesting of the new configuration. Examples of such changes are choice of parts, such as de-rating factors, workmanship levels, manufacturing processes, such as certain fluxes or potting recipes, and surveillance, such as witness plates for cleanliness.

The implications may extend beyond engineering or manufacturing. They may also include requirements on the space flight project for applying oversight or insight measures in the technology lab. In summary, the technology project will be expected to comply with a host of potentially restrictive requirements that may increase cost and reduce flexibility.

22.3.3.5 Define 'Mission Success' at the Beginning

Project managers can easily believe they are the keepers of the definition of mission success, but this simply is not true. So they must ask for the stakeholders' definition of success and then universally apply it throughout the project's life cycle.

22.3.3.6 Bridge the Needs of Stakeholders and Technologists

Technologists want to fly their technology, and stakeholders want their data. That is why both are in the business. In many cases, they will not even know each other unless a scientific measurement, engineering assessment, or other mandate has driven the technology's development. For these situations, project managers are the bridge between both parties and may need to take one side or the other in order to converge on the best overall position for the mission. When this happens, technologists may feel that the project manager is not acting as their advocate. In an important way, that is true! Project managers should advocate for the space flight project, not for a technology. Remembering this, especially during difficult times, will make it easier to frame conversations in the most constructive terms.

22.3.4 Shaping the Integrated Mission Schedule

A common thread through all of a project's life cycle phases is the schedule. It is the flight plan for any project but especially for space flight missions, in which many parallel activities take place. Because schedule is so critical, project managers and many stakeholders pay a lot of attention to it.

Very few projects are built from the top down, because this approach is fraught with risk. Rather, scheduling almost always takes a bottom-up approach: project elements develop their own schedules, often getting coordination or support from a dedicated scheduler on the space flight project. At specific intervals, each element submits its latest draft version of the schedule. All schedules then consolidate into a single, main integrated mission schedule (IMS).

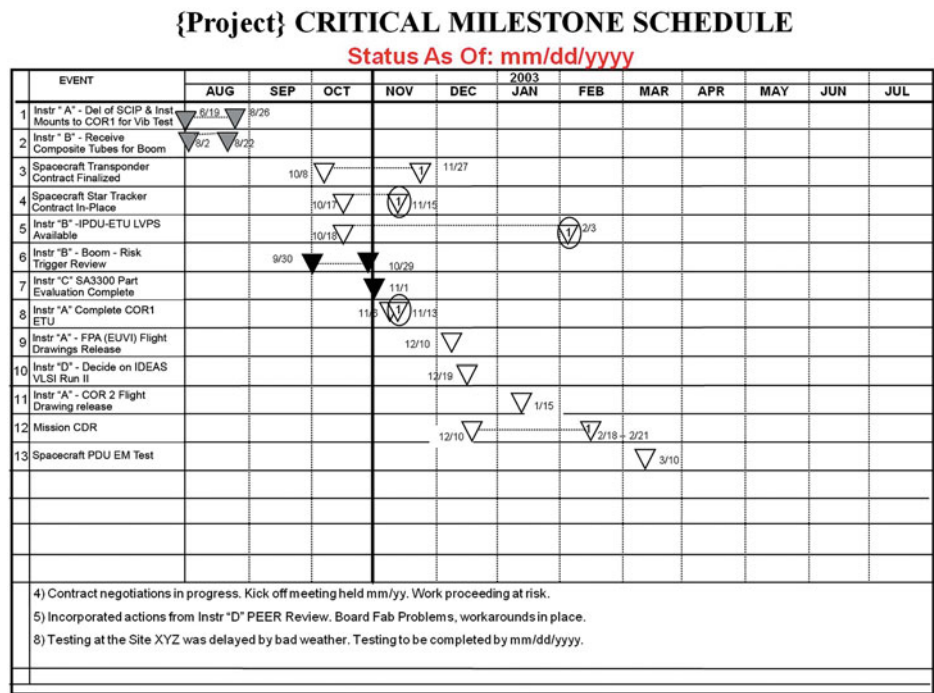
As space mission developers, we should not underestimate the time and effort needed to create an element-level schedule or the number of iterations we may need to incorporate required adjustments. These adjustments might be necessary to more efficiently order the activities, synchronize predecessor and successor tasks, best use the workforce, or create a baseline IMS. All of this effort occurs before the work has started, while the project is tracking schedule performance through milestone completion, the earned value schedule-performance-index (SPI), or similar tools for managing schedules.

22.3.4.1 Tailor the Schedule

Creating a schedule for a reasonably complex project activity is specialized, so it often devolves to professional schedulers, but project managers and technologists still need to understand basic scheduling principles. Among concepts and calculations that should be familiar to any project manager are critical path, margin, reserve, forward and backward calculations of slack, and dependencies. We will not repeat here what you can find on scheduling in other project-management textbooks. However, we can add to the conversation by talking about how project managers might tailor an IMS and why that is important.

In the life cycle model used by NASA, DoD, and many private companies in the United States, project managers and teams do not decide on the transition from one phase to another. Depending on its criticality, the hosting organization or even first-tier stakeholders also may not make this decision. Instead, this action goes to a decision authority: a senior-level person who is advised by a board or council on the project's maturity. The decision will depend on the results of auditing necessary work products and similar factors. Only this authority's express authorization can

Fig. 22.6 Critical milestone schedule. These schedules are an excellent format for focusing on major activities that can strongly affect project performance. By reviewing critical milestones the project managers can efficiently align the project’s focus



transition the space flight project into the next phase. These points in the project are called key decision points and they drive all aspects of the space flight project’s (and the technologist’s) activities because they represent the end of work that all project participants must complete at a certain point in the schedule.

Key decision points (KDP), i.e. milestones, are points of zero duration intended to mark the moment when work starts, finishes, or requires a key decision. For a space flight project at NASA the most important of these are the key decision points. Besides KDP milestones, an organization is likely to require the project manager to insert internal milestones into the schedule for such activities as budget reviews, workforce actions, pre-briefings for upcoming meetings, and reporting requirements. As described earlier, stakeholders set the policies, standards, and guidelines that constrain the project. They also must become milestones in the IMS, where they will fold into the technology project’s activities. Examples include peer reviews, formal independent reviews, and meetings with external groups.

Project managers will insert more schedule milestones at their discretion to benefit the project. These milestones may be project-wide or specific to a group. They may involve an activity, or simply be intended to synchronize activities across the project. For example, the manager may convene an all-hands meeting to review the lessons learned after completing a key activity.

When necessary, project managers will define more milestones for an activity that is running into difficulty. These may be as simple as timed meetings to review status, progress assessments, or—in unfortunate circumstances—

difficult decision points, such as substituting a new technical approach when the baseline approach isn’t working.

22.3.4.2 Manage Schedule Performance

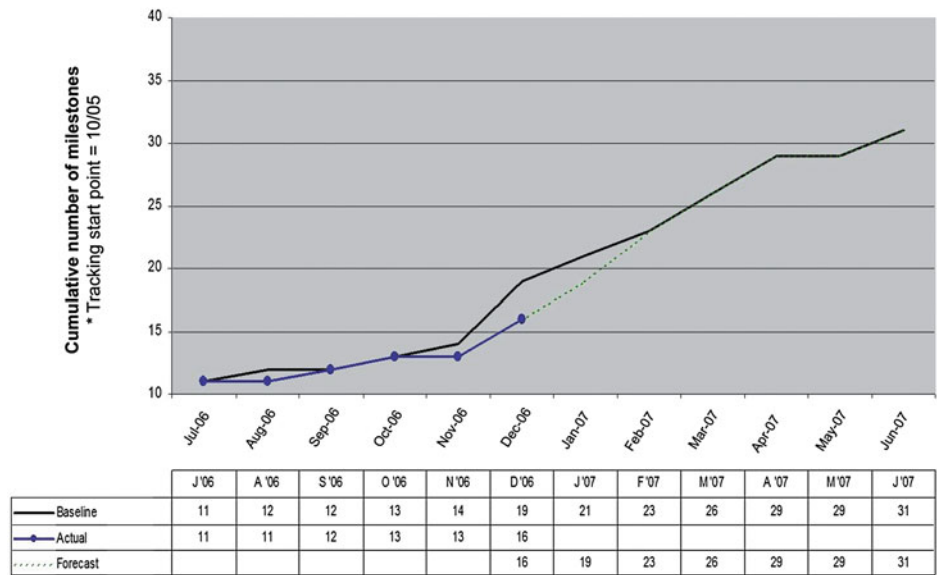
Most space flight projects have a huge number of schedule items in the database: 2,500–3,500 is not unreasonable. However, trying to display all these entries would be unreasonable. As a result, a project rolls up its IMS into a top-level schedule that is easier to read. It presents large relationships, including the project’s critical path—the sequence of activities that has zero or the lowest slack. A top-level schedule that GSFC would use is illustrated in Fig. 22.5.

The top-level schedule is an excellent way to summarize the IMS and inform project members, stakeholders, and the organization, but it is not the best format for managing schedule performance. That requires a critical milestone schedule (see Fig. 22.6), in which specific activities are drawn out of the IMS database and displayed in a rolling time-window format.

Project managers and teams can use a critical milestone schedule to cut through myriad data and see what really affects the project’s performance. By reviewing these focused formats monthly as part of project meetings or regular business processes, they can draw attention to important activities that are late, coming soon, or represent workload over a given period.

When it comes to managing projects, schedule data offers project managers and technologists a lot of flexibility. As an example, despite the best efforts to plan work accurately, one or more parts of a project will always take longer

Fig. 22.7 Cumulative milestone tracking schedule. Project managers can use schedule data in special formats to focus on critical work and its completion rate



than planned. Also, certain tasks will have many deliverables, such as engineering drawings, that need to be done by a certain time in order to meet a schedule milestone. Lastly, it is also probable that one or more activities within the project were incorrectly planned as a result of missing information, unanticipated risks, or human error, and therefore requires replanning. In these instances, putting schedule data in a special format allows focused tracking of the effort. An example of this format is cumulative milestone tracking. Its power lies in its flexibility. We can use it to track project activities within specified periods, for specified schedule items, or any combination of the two. Figure 22.7 illustrates how this format helps us define the plan for completing work and then tracking progress against that plan.

With focus on the IMS and insights from the schedule database, we may discover that a technology development will take more time to complete than even the most pessimistic scheduler assumed. This problem arises most often when a flight mission adopts an immature technology, but even with technologies developed in advance to meet mission requirements, we may still reach a point where no revision is possible within the program’s constraints. The IMS will be an excellent early indicator of this problem, as long as a team does not ignore or discount it because they believe their original decision was the right one. The key symptom for this problem is a continual need for more time, which at some point the space flight project will not be able to grant.

One way to mitigate this risk lies in improving detail during the interaction that takes place between the space flight and technology projects early in the life cycle phases. Normally, this interaction occurs before anyone commits to incorporating the technology into the mission baseline. Schedule considerations may therefore be but one way to

determine whether the technology will work for this application. If the technologist can provide schedules for the technology-development activities, the project can make fewer assumptions and plan better. The goal here is to reduce the amount of time a technologist needs to plan new activities. If schedule inputs aren’t available, the technologist must plan sufficient time to enable both sides to pull this information together, define special schedule formats early, and then convene regular meetings to assess performance against the defined plan.

22.3.4.3 Handle Tough Scheduling Decisions

Ultimately, a technologist will know whether a project is meeting its initial commitments. After missing significant milestones, the technologist and project manager will have recognized the problem and discussed the negative performance trend. Unless things change, the project manager faces a tough decision. Nobody enters into an agreement intending to fail and, understandably, nobody wants to have his or her technology removed from a space flight project. But it does happen when the project can’t tolerate schedule delays any longer if it is to support the mission paradigm. Planetary missions are an excellent example: launch windows are limited, so delays are costly.

How do we determine when this action of last resort is necessary? Although many approaches are possible, one of the fairest is for the technologist and project manager to define the key milestones at which, if the trend has not changed, the technology must go. By negotiating, both sides can work together to provide a technology team the maximum time while still protecting the space flight project.

This is exactly what happened to an instrument on a heliophysics mission at GSFC. The instrument’s original design took advantage of a new detector that was unproven

in space but showed great promise in the laboratory. The space flight project incorporated this instrument into its payload with the new detectors, and the integrated team began building the instrument. Unfortunately, they couldn't mature the detectors in time, and the instrument's development fell significantly behind schedule. Working together, the project manager and technologist developed a recovery plan—defining clear milestones and what was needed at each one. They also teamed to brief the stakeholders. The recovery plan allowed the technology project several months to recover, but they were unsuccessful. The project replaced the instrument's unproven detectors with more mature ones, and the mission eventually flew into space and is still operating as of 2013. This was a perfect example of a strong collaboration between the technologist and project manager giving the new technology every chance to fly but, once it clearly could not overcome the obstacles, agreeing that the space flight project should continue its march toward launching on its own.

22.3.5 Defining, Phasing, and Managing the Budget

Accurately defining a project's budget requirements for the entire life cycle is the key to any successful project. This task becomes more complicated as the project schedule grows longer, perhaps spanning several fiscal years, or becomes more complex. Even in projects with a well-defined budget, failure to manage expenditures sufficiently closely and identify trends early on is a recipe for failure.

Regardless of the cause, a poorly defined or managed budget will at best prematurely drain project reserves: the relatively small amount of extra money available to the project manager. In extreme cases, not budgeting realistically may mean a project fails to satisfy stakeholder commitments, thus delaying its transition into subsequent life cycle phases or even being canceled. Space flight projects recognize the importance of continually managing their budgets.

The mission paradigm for a space flight project will largely dictate the tone set by the project manager and business team. This means technologists must anticipate different priorities from mission to mission and engage the project manager early in the relationship in order to confirm them. Is this a technology-demonstration mission in which meeting functional requirements is paramount? Or, is this an operational mission for which we must deliver at least a minimum level of performance in order to satisfy the schedule? Once all project participants discuss and understand the mission paradigm, they can begin to define the budget.

22.3.5.1 Estimate Project Costs

The goal of cost estimation is defining accurately the resources required to complete a task, but that can be extremely complicated and challenging. For space flight projects, it is downright hard and—more often than not—the resulting estimate under-predicts what the project will require in order to be successful. So planners pay a lot of attention to cost estimating and often continue it throughout the project's life cycle.

The space flight project's business team will take the results of the engineering process to define the mission and develop a basis of estimate (BOE) for the team. The highest-ranking resources manager on the project normally handles the BOE process. This manager's team typically creates the BOE materials, outlines the schedule, and establishes ground rules. The BOE materials usually include two important pieces of information: budgeting guidelines and the budget-definition template.

Budgeting guidelines offer specifics on the process the team will follow and may include such topics as milestones for the process within the project, reconciliation periods, direction for funded and unfunded contingencies, workforce skill categories, their associated rates, and escalation factors, and direction for handling indirect costs such as fees, administrative expenses, and travel.

The cost account manager (CAM) normally completes the BOE templates—in the best case, with input from the technical team working with the CAM. The CAM provides the requested information and is encouraged to identify any assumptions outside those that the project captures, as well as any risks that may cause the plan to change.

With project-wide guidance in place, the technology project may begin to develop their cost estimate. How they do this will depend on the maturity of the technology being applied, the amount and complexity of changes required for the space flight project's application, and the team's experience under similar conditions.

Many schedule and cost-estimation tools are available to support budget creation. No single tool has been able to generate an accurate budget in all situations, so technologists and their teams—consulting with the space flight project's financial and business community—must determine the tools that will drive production of the BOE. Table 22.6 summarizes many common approaches and their best use.

22.3.5.2 Estimate Workforce Costs

The project team will have a staffing profile to cover the time when work is scheduled for completion. The major variables in defining costs are the skill levels of the people who work on the project, how long they will be charged to the project's budget and the total number of workers for each monthly budgeted period. The project manager will negotiate with the technologist concerning the duration and staffing needed

Table 22.6 Cost estimation tools

Approximate accuracy	Tool title	Remarks
Coarse estimate (lowest accuracy)	WAG, also known as V-ROM	Least accurate of all the techniques: tantamount to a ‘guess’ Fastest, cheapest estimate
	Rough order estimate (ROM)	Typically defined by a stakeholder or the project manager Marginally better than a WAG (V-ROM—sponsors can expect accuracy no better than $\pm 50\%$)
Coarse estimate (lower accuracy)	Delphi	Best used on one-of-a-kind build where we have little or no experience or data Uses experience of one or more subject matter experts to hone in on an estimate May use global knowledge of the subject matter experts or be anonymous
	Top-down	A team-based technique for which management predefines allocations across the participants May use under competition with capped life cycle cost A tempting way to define costs but flawed because it separates the estimate’s creator from the actual producer of the work
Coarse estimate (low accuracy)	Standards	Organization-based and typically used for parts of the project—for example, management equals 10 % of technical hours
Analytical estimate (intermediate accuracy)	Extrapolation	Uses historical data to estimate price when present activity is out-of-family (desired data point is outside the range) May use with the Delphi method to bound uncertainty
	Analogy	May use when present build is in-family with past activities (desired point is inside the range)
	Parametric cost	Algebraically represents cost using variables common to all builds, resulting in cost estimation functions Uses past performance to fine-tune coefficients and address historical risks
Specialist estimate (higher accuracy)	Pricing team	Employs specialists in CE to help the project develop an estimate Typically, holders of the most accurate database of historic performance Most proficient at using CE tools and customizing the combination of tools for a specific instance
		Team-based estimate (highest accuracy)

All parties in the space flight mission must estimate the costs for their tasks. Although no one tool accurately does this, they can use many available tools

to tailor the technology in the lab and to integrate parts of that team with the project team. How they interact will depend on the work inherent to a particular life cycle phase. Typically, the first iteration of workforce planning produces abnormal peaks and valleys in the profile. Because people are involved, these peaks and valleys are not acceptable or efficient. So the project team will undergo ‘resource leveling’ to reorder the work and smooth out the profile, even if it slightly lengthens the overall schedule.

Technologists must not assume the project will ‘carry’ their lab or many of their people. Project managers negotiate the right level of support, at the right time in the life cycle, and for the right duration. Technologists must recognize the importance of negotiating a valid plan and then arrange the lab’s affairs so that it can happen.

22.3.5.3 Estimate Task Contingencies and Project Reserve

No one can accurately define the future, but project managers are highly skilled in setting aside funds to address uncertainties. Technologists also are empowered to identify where uncertainty exists and where they cannot adequately control external influences. Classic examples are delivery of long-lead parts or availability of integration and test facilities. In these instances, a technologist and the project team will add contingency and then track it very closely because all players in the project see the consumption of contingency as a measure of potential problems.

Space flight projects won’t recognize all uncertainties in the budget estimate. Even when we believe our plans are accurate and complete, actual performance may differ,

resulting in the need for more funding. Project management will respond by directing the resources community to hold part of the money in reserve and then release it to make the necessary adjustments.

Technology and space flight projects are inherently difficult, so no project's performance is perfect and nor are all plans right. Therefore, technologists and project managers must carefully monitor a technology project's performance, just as they would for all parts of the space flight project. The key is to ensure they use reserve money for transient conditions and not for systemic flaws. They cannot allow such flaws to continue in any part of the project.

Project managers and technologists should define how they can lessen a project's scope—called descoping—while emotions are in check. The best time is early in the life cycle, with requirements defined (though minor work may continue), the technology development plan in place, and the schedule and budget taking shape. By definition, descoping options are clearly defined, managed reductions of capabilities that reduce the work needed while still ensuring the mission will meet performance requirements. They are an important arrow in the project manager's quiver because taking descopes can be a last resort to keep project cost within stakeholder constraints, reduce drains on project reserve, and protect against more schedule erosion. The challenge is that people never want to produce something less than what they envision, so descoping discussions can become heated. Hence, managers should make descopes part of planning, control their configuration, and apply them only when necessary.

22.3.5.4 Repeat the Budget Process

Because funding is the life's blood for a project, and sponsors often have several projects running simultaneously at different phases of the life cycle, technologists can expect a detailed annual review of the budget. Typically in the middle of the fiscal year, each space flight project, 'marks' the budget by using actual costs for all prior years and the first 6 months of the current fiscal year, and requires an updated basis for estimating the remaining 6 months of that fiscal year and all future years left in the life cycle.

This information feeds into an updated project budget, which may be very different because of internal adjustments needed to cover poor performance, increased costs, or change requirements. Also possible are external adjustments, such as sponsor-specified budget reductions, lack of a launch vehicle, or loss of expected institutional funds. Managers must reconcile the budget to ensure life cycle cost commitments are intact, the revised budget is fully defensible, and the reserve levels cover the mission's remaining life cycle phases. For obvious reasons, technologists are involved directly in this annual budget process.

A lot of information comes out of the budget process beyond knowing how much time and resources a task is planning to use. The estimate at completion (EAC) is one of the most important. This number projects what the final cost will be for any task, series of tasks, or the entire project. We can calculate the EAC in several ways, and not all algorithms are meant for all situations. Are unique past challenges finished? Is the difficult work about to begin? Are process changes available to increase schedule efficiency? These represent some of the questions we need to answer in order to project an accurate EAC. The technologist and his or her business team must provide as much information as possible to the space flight project's business team to enable the latter to calculate an accurate EAC. This value will be a major driver of how the project manager handles this aspect of the project during the next budgeted period.

Money constraints are a fact of life for most organizations. Although they will continue to have a portfolio of missions spanning all types, these constraints are likely to be strong enough to lower their tolerance of schedule delays and cost overruns. Therefore, the stakeholder's actual tolerance of risk determines the accuracy of estimates and candor in reporting.

22.3.6 Managing Risk Continuously

All common threads are essential to a properly working project, but only a select few are both highly visible and hard to apply correctly. Continuous risk management (CRM) is one of these 'special' common threads. Therefore, technologists and everyone working on the technology project must fully understand the risk process, become skilled at applying it, and calibrate it accurately to the space flight project's cultural norms during all project phases.

22.3.6.1 What is CRM?

CRM is the process of looking across all aspects of a project to identify early on those issues that may go wrong. Because successful CRM identifies the problem before it causes harm, stakeholders and organizations have embraced it fully in an attempt to save budget, save schedule, and increase the overall likelihood of mission success. In fact, many organizations mandate its use. Figure 22.8 shows the iterative process GSFC uses in response to NASA Headquarters policy [15]. Although CRM may be documented at the top level, the space flight project will know its internal process, the format for reporting results, and how to use identified risks for the budget process.

Project managers can lead CRM for small projects, but larger, more complex projects usually hire a risk manager. In any case, project managers must be fully involved in the

process. If it is unclear whether hiring a dedicated risk manager is necessary, we recommend hiring one because of high demands on the project manager's time and the tendency to underestimate risk. If budget is an issue but hiring a risk manager makes sense, the project manager should define a staffing profile with more effort at the start and then enter an internal-project CRM milestone in the integrated mission schedule. At this milestone the risk manager and project manager review the CRM activity to date and, with a much better understanding of the project environment, determine how best to proceed with the support of a dedicated risk manager. This conversation often is easier than expected because proficient risk managers are not interested in working on a project with too little workload to keep them fully engaged. When selecting a CRM practitioner, keep in mind that certification programs are now available; the U.S. Project Management Institute (PMI) is one example.

22.3.6.2 Why is CRM so Important?

With all the literature available on CRM and the number of organizations that use it, we will assume everyone on a technology or space flight project has been exposed to it. Here, we are concerned less about how to use CRM than why it is so important.

Risk management is more than a tool, it's a mindset. Once we become skilled at thinking in risk terms, we tend to do it all the time and it becomes second nature and pervades our work. Whether it is an engineering, budget, or schedule challenge, we include risk in the analysis and try to make the plan more robust. As a result, CRM specialists tend to generate products of higher quality, at or below the negotiated cost and within deadlines. However, project managers and technologists cannot assume that all members of their teams have this mindset, so they need to ensure that everyone is thinking about CRM. They can best do so by demonstrating the process publicly, underscoring its importance at general meetings, and calling attention to actions that reduce the project's overall risk exposure.

No matter how much effort we expend, no plan is perfect. Despite the project team's best efforts, every plan will encounter unidentified risks. Therefore, nobody should assume that CRM will enable their project to escape them. We want to use CRM in order to reduce risk sources to two categories, so that we can better manage net inherent risk.

- *Known unknowns*—a source of risk because we lack knowledge in an area where we know the information exists. The project will uncover this information with more effort or time, and CRM will then continue as normal.
- *Unknown unknowns*—a source of risk because we lack knowledge in an area where the information will not become available. We commonly refer to these issues as

part of 'Murphy's Law' and have to deal with them as they arise.

22.3.6.3 How Should Projects Manage Risk?

Most importantly, the project alone cannot or should not answer these questions; managers must discuss risk with the stakeholders and organization. The project manager will then calibrate the project to this level of risk tolerance. At NASA, risk tolerance for space flight missions is an important conversation—so important that official policy and procedures have codified it [16]. How a space flight mission deals with risk depends on the mission paradigm, the stakeholders' tolerance, the life cycle phase, and how a project assesses the consequences if the risk event occurs.

Implications from the mission paradigm and stakeholders' tolerances are self-explanatory, but those from the life cycle phase may be less obvious.

Often, managers think of CRM in terms of risks uncovered during implementation, but this orientation limits the tool's power. If risk management is ingrained in a project, it will identify and factor risk into the earliest mission definition. The cumulative effect of all accepted risks is the 'risk posture' at the start of the project. Stakeholders review this risk posture as much as the mission itself when someone is trying to sell the mission to them, especially for a competitive environment in which mission cost typically is capped. This early stage in assessing and accepting risk is a critical time for the project and the technology team.

At the start of the life cycle, the project team is synthesizing the baseline plan of what to build and how to build it. The space flight project is asking itself whether the inherent risks of a new technology are appropriate. Technologists who can explain the benefits of their technology in engineering and risk terms can best convince the project to apply it.

Even when technologists and project managers fully discuss why to use a new technology, sometimes it is appropriate for the latter to decline. This decision should result primarily from analyzing the technology's ability to satisfy the project's requirements. In other words, the technology may be fine in and of itself, but another solution may represent a lower cumulative risk.

Once the project establishes a mission baseline including the new technology, it has an initial risk posture. From this point forward, the technology and space flight projects transition into the traditional mode of identifying, classifying, and mitigating risk. They will use a documented process that fully incorporates the technologist and technology project. As the project manager identifies a risk, he or she may require the technology project to analyze it, determine its trigger, establish its consequences, and propose a possible way to mitigate it. These studies may not be

funded, so technologists must recognize their implications for the overall schedule—especially their cumulative effect.

Not to fly is better than to suffer a bad flight. We have no cost-effective remedy for matching a good technology with a bad application. Large or small risk entries in the database will not make the wrong application of any technology right. Because everyone’s reputation is always at stake, we need to use higher-order thinking and admit facts when the match is not right. If the mission’s purpose is to fly the technology, then the technologist and project manager must work together carefully to produce an accurate, risk-based decision for proceeding.

Opportunity management demands attention too. CRM focuses on the negative, but what about good changes to the project’s plans? Called ‘opportunity management’, this possibility gets discussed in the academic or standards communities, but it doesn’t get the same level of attention from projects or their stakeholders. The reasons why are numerous and varied. In our experience, the main reason is the naturally occurring bias that exists in almost all projects when planned. Meaningful positive events are less probable than the risks we have all come to anticipate. In other words, when talking about Murphy’s Law nobody ever says Murphy may have a sibling! Regardless of the reason, technologists and project managers must discuss at risk meetings those future events that could end sooner, cost less to complete, or reduce requirements—thereby lowering project costs or the overall residual risk.

22.3.6.4 How Do We Report Project Risks Effectively?

With CRM being a critical activity within the project, it is not surprising that reporting risk is an equally important activity.

The goal of CRM is to identify risks and then evaluate each one in terms of the likelihood and potential consequences of its occurrence. The process is not supposed to be limited to any one aspect of the project; it is a comprehensive tool for a repeatable process that looks at technical, program, and other key aspects of the project’s execution. To categorize each risk, the project uses a ‘key’. This helps to keep an individual project’s evaluation consistent and gives the organization a unique perspective on risks common to multiple projects when applying the same key to all. Figure 22.9 shows an example of the risk-evaluation key that GSFC uses across our project portfolio.

As risk is so important, a large number of stakeholders want to know the status of a project’s CRM, so we normally summarize risk information in standard reports. Figure 22.10 shows an example of a project’s summary risk chart. Most projects have many risks, so reports typically focus the audience’s attention on the top ten.

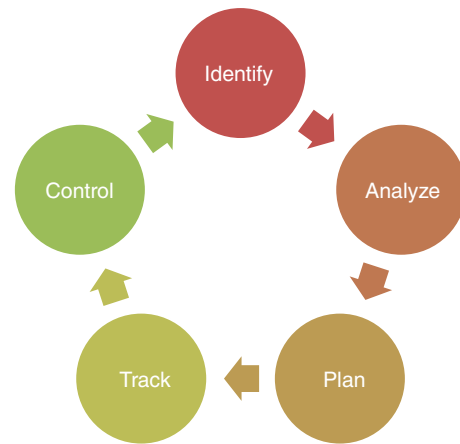


Fig. 22.8 Process for continuous risk management. The key to risk management is finding problems before they occur. It repeats throughout the project’s life cycle

Although it is important to understand a project’s overall risk posture at all times, summary data seldom meets a stakeholder’s need for understanding a specific risk. So the risk section of a project report usually pairs the summary chart with a risk focus chart that offers more insight into the top ten risks. The normal practice at GSFC is to provide this additional focus only on ‘red’ and ‘yellow’ risks. Table 22.7 shows what additional information is provided for each of the risks we’re focusing on.

22.3.7 Managing Team Activities

Project managers get work done through people, so they focus most of their time on the people critical to a project’s success. Our experience from many projects has identified some best practices for project managers of a space flight project working with a technology project: project meetings, configuration change control, and management information systems.

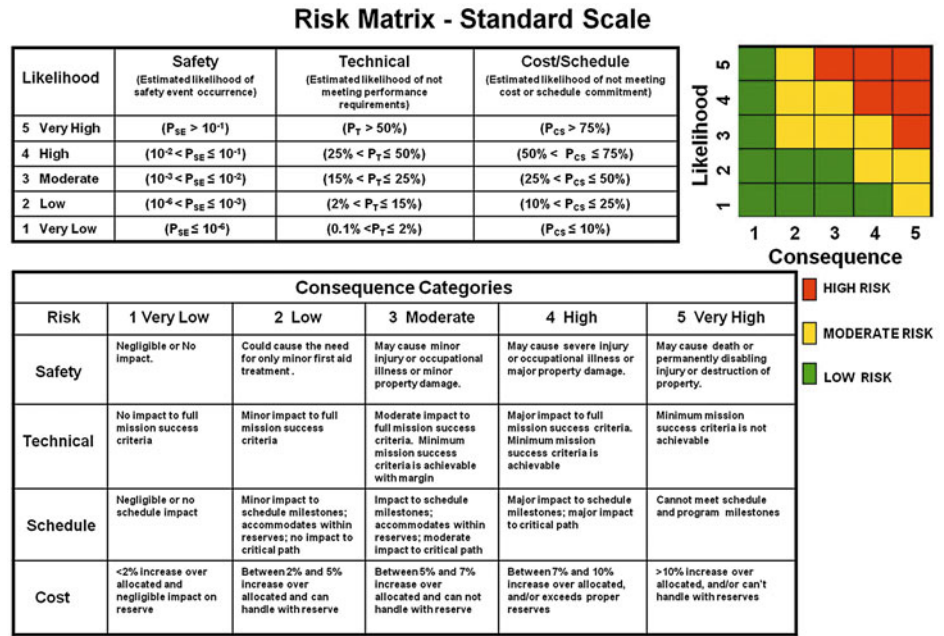
22.3.7.1 Manage Project Meetings

Communication is a key to mission success: with stakeholders, within the organization, and especially among project team members. Regular meetings for the entire project team—space flight project and selected technology project people—are a key way to ensure everyone understands the project’s current status, knows the detailed plans for upcoming near-term tasks, and has a clear vision for the challenges that lie ahead.

If geographical separation, language barriers, or other complicating factors are at play, project managers need to communicate consistently with the project team.

Technologists and space flight project managers must ensure their respective teams communicate well among

Fig. 22.9 Risk item evaluation key. For large projects, many different people evaluate risks. To ensure consistency in risk categories, project members commonly use a ‘key’



themselves. However, whenever the technology is being integrated into the space flight mission, both managers must meet the extra challenge of linking their approaches effectively.

Senior staff meetings. Accomplished project managers know they cannot do everything on their own. That is why projects have deputies, group leads, and other senior-level positions. The more closely this small team aligns with the project manager’s priorities, the more integrated the entire project will be. Thus, experienced project managers add a small, dedicated meeting to their weekly schedule that brings together the senior staff: typically, the deputy project manager, deputy project manager/resources, mission systems engineer, instrument systems manager, observatory manager, and ground segment manager. The purpose of this meeting is to address sensitive matters that would be inappropriate for a project-wide discussion.

The senior staff work best in a short meeting, so each participant gets a limited chance to talk. For efficiency, participants should come to the meeting with their topics already identified. Although sharing status is always important, the main purpose of this meeting is to focus on key challenges facing the project. To establish the meeting as a cultural norm, project managers should put it on the calendar for every week of the year, canceling only when necessary.

Technologists also should have their own senior staff meeting scheduled in parallel, so that they stay current on issues that affect the technology project and are likely to become part of a conversation with the project manager. If a challenge needs a consolidated team’s attention, either manager should feel empowered to request a joint meeting.

That would give key decision makers the chance to discuss issues candidly and define appropriate action plans.

Weekly meetings for project staff. Regular meetings are important to convey project information to the entire project team: space flight and selected technology project people. They ensure that everyone understands team status and plans for near-term tasks, and project priorities. Project people are aware of the environment in which the project is developing, but they do not have the project manager’s insights on current issues. The weekly project staff meeting is an excellent way to convey how issues affect the project and the team’s response.

All-hands meetings. When a project involves many people, or parts of the project are separated because of geography or other factors, weekly project staff meetings with all members of the team may become impractical. However, project members who do not participate in a weekly meeting still want to hear from the project manager. All-hands meetings scheduled periodically throughout the year can satisfy this need. Whenever possible, project managers should visit the team and lead the all-hands meeting in person. However, remote technology can substitute for personal contact during times of fast change or project execution.

22.3.7.2 Manage Change

Early in the project life cycle the team baselines final decisions for carrying out the mission in its project plans, specifications, interface-control documents, integration and test plans, operational procedures, integrated mission schedule, budget, and risk plan. All of these documents are subject to configuration control. Without universally applied

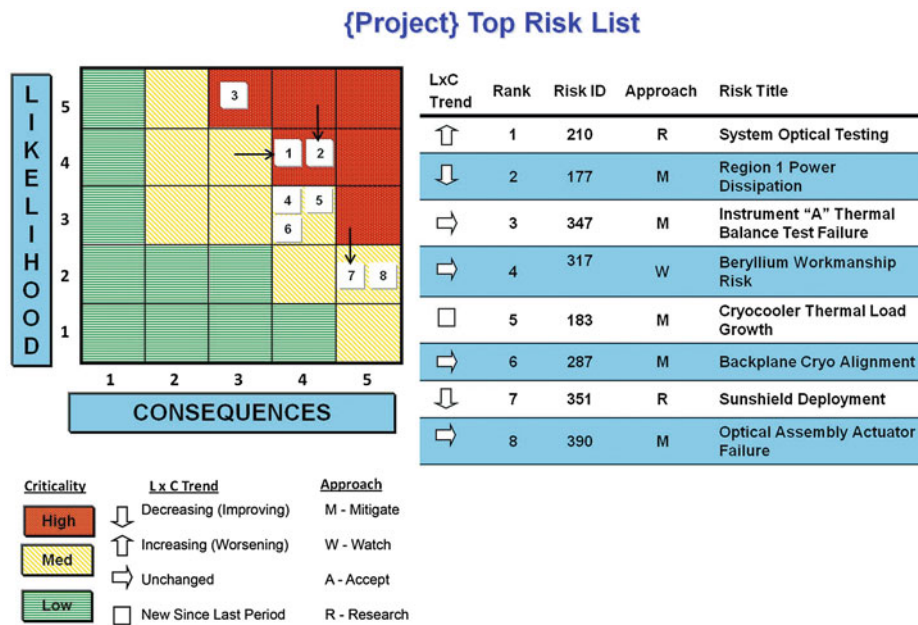


Fig. 22.10 Summary risk chart. Projects normally have many risks, so reporting typically focuses on the top ten risks for the project

Table 22.7 Focused reporting of critical risks

Reporting element	Content
Header and footer	Lists project name and related identifying information If the risk focus information is not part of a larger report, the header and footer documents the date and other versioning information
Information on specific risks	Provides the risk's title, rank, current trend, and expected closure or acceptance date
Risk ID	Unique ID number for each risk when it first goes into the risk database. Permits mapping of this information to any previous discussion on this risk; especially important if the risk's focus changes in a minor way over time
Risk statement	During its evaluation we characterize risk in a construct the organization and stakeholders agree on We express risks in one of two ways: IF {the following occurs...}—THEN {the consequence(s) will be...} GIVEN THAT {the following can occur...}—THERE IS THE POSSIBILITY {that the following consequence may occur...}
Approach	Repeat the project's approach to addressing the risk: research, watch, monitor, or mitigate If this approach has changed since the last report, explain the new approach's rationale
Status	Because the focus is on a risk that can strongly affect the project, explain the team's current efforts to prevent it

Stakeholders want more insight into risks that have the greatest effect on the project if they occur. Complete information builds stakeholder confidence in the project's risk process

change-control procedures, the project will be at very high risk of an error or even failure. This concept of change control is so important that the organization usually defines the policies—and sometimes the specific work instructions—that all projects must follow. In some cases, organizations adopt standards documented by external groups, such as the International Organization for Standardization.

Some change is inevitable, but not all change is necessary. To decide which changes to make, the space flight project manager convenes a change control board (CCB).

Large, complex projects may use a hierarchy of boards. They may assign lower boards responsibility for controlling the changes that affect only their project segment or that cost less than a preset amount of budget. An example would be the GSFC's Earth Observing System Data Information System (ESDIS). This project has three CCBs: overall project, flight segment, and science segment. However, board members need discipline because changes that exceed a lower board's authority must go to the project-level CCB for disposition. Success depends on fully coordinating the

change across all affected parties, updating documentation to reflect it, annotating budget cost plans to capture the new plan, and focusing risk on the new baseline to ensure the project has new mitigation plans.

The technology project's requirements are integrated with others in the space flight project, after which all go under configuration change control. As challenges arise in the technology effort, the technology team must exhibit the same level of discipline as the rest of the project by submitting appropriate change-control documents. If the two projects have strong collaboration and communication, this process will work efficiently.

Whenever different organizations host the technology and space flight projects, each organization has its own defined process and cultural norms for how to change the mission baseline. At times, this can be a source of frustration between technologists and the project managers. The very essence of change control is to limit the number of changes, but technologists and their teams live in a world where they must explore ideas in order to improve and advance. This tension, though natural, must be managed to keep deliveries flowing smoothly to the project. At the same time, technologists must oversee their laboratory procedures to minimize changes in delivery items.

22.3.7.3 Use Management Information Systems

As society envisions larger, more complex systems, the projects chartered to deliver these systems must develop new approaches to carry out project functions, minimize cost, deliver in less time, and maintain the same level of residual risk. That means new challenges in storing, distributing, and protecting data. One way organizations have evolved to meet these challenges is by using computer-based applications for library functions, electronic mail, reporting, and performance monitoring.

We can collect these applications for automating project functions into management information systems (MIS). Technologists may expect their project to be included in the deployment plan for these applications, but they must work with project managers to define data-protection measures before the MIS goes operational.

Reporting is a multifaceted topic that warrants special attention, especially for missions employing new technologies. Reporting activities command a lot of the project manager's and technologist's time. They must coordinate to report general and specialized requirements completely.

22.3.7.4 Meet Many Reporting Requirements for Space Flight Projects

Space flight missions have many stakeholders whose diverse interests will span aspects of the overall project, technologies, and budget. As a result, many people want to

know how the project is progressing, what challenges it is facing, and what the projections are for final cost, delivery date, and capabilities. Several reports are necessary to satisfy these interests.

Weekly activity reports. Most organizations require a report each week, in a format typically defined by the requesting organization. By adhering to this format, the space flight project supports organization staff who combine several reports into a larger, more general weekly report. Consequently, project managers and technologists need to see themselves as part of a larger community who report status. Table 22.8 offers a basic format we may use to report weekly status if we do not have a defined format.

Monthly status reports. Most organizations also require space flight projects to deliver a monthly status report. The project manager, or in his or her absence another senior project manager, normally makes this formal report to a group of stakeholders who may or may not be advocates for the project. It contains more complete status than the weekly report and conveys a complete snapshot of the project at this one point in time. The information spans all aspects of the space flight project, including technical, schedule, and budget performance. The report also incorporates updates to issues, risks, and budget projections, such as the results of earned value management.

In some organizations, project people meet with first-tier line managers to provide a real-time summary of what they will present. This approach adds the benefit of giving managers a chance to identify crosscutting issues that are best addressed by management above the project manager. Because the monthly status report is formal and complete, organizations typically predefine its format and share it with all reporting sources in the project. Table 22.9 shows the content of a typical monthly status report at GSFC.

Miscellaneous status reports. Nothing draws more attention to a space flight project than poor performance. The definition of poor performance can range from a real inability to progress within budget or schedule commitments, to unexpected test results, to frequent changes in requirements. It also can range from a perceived inability to communicate clearly and consistently with stakeholders, maintain the desired attrition for project people, or negotiate effectively with partners or vendors. Yet, space flight projects that are not poor performers may still receive excessive interest because the mission is important to the organization's portfolio, it carries special subsystems (such as nuclear power subsystems), or the political environment imposes pressure on the outcomes.

As a result, project managers may expect more planned and unplanned requests for status information in a format negotiated with the requestor or organization. Technologists should be sensitive to these special factors and participate fully, as appropriate.

Table 22.8 Generic format for a weekly report

Weekly report section	Content
File name	Give each file a unique time-based name that conveys the weekly report. Formats vary, but a common approach to help us sort several weeks of reporting data would be: YYYYMMDD_{Project Name}_Weekly_Rpt.{ext}
Report header	Use modern word-processing applications to create a header that repeats on each page. Although we may not be able to see this header on the screen, it's always visible when printed Include top-level information, such as the project's name and the period covered in the report As a vendor or support contractor, you may include the contract number and current period of performance
Report title	On the first page, provide the reporting group's name As a vendor or a support contractor, you also may include the task number for the reported work. This task number would correspond to the contract number in the header
Highlights	In bullet form, summarize key accomplishments during the reporting period Need not complete all items. In some cases the highlight may show you're starting work, continuing a task, or offering insight into an activity's progress. The number of bullets isn't important, but covering the work is For larger projects involving many segments or activities, can subdivide highlights section by segment For larger or critically important projects, use an executive summary at the start of this section, so readers will have context for the rest of the report
Issues or concerns	'Concerns' are challenges you are still working but the reader should be aware of. Doesn't request management action 'Issues' are challenges you cannot solve on your own; management attention is required or the impact is expected to occur Balance the stakeholders' need to know with your understandable need for a chance to solve the problem. Carefully determine the level of 'surprise' that may result from the impact for a challenge you never reported
Risks	Summarize key risks facing the activity but don't repeat all the risks in the risk list
Key meetings	List important meetings you attended during the reporting period List important meetings coming up within, say, the next 4 weeks
Milestones, schedule, and deliverables	Include a table summarizing the pending milestones and deliverables. Include a scheduled completion date for each one and a 'status' indicator summarizing information the reader should take away. Examples are completed, started, or on-hold pending resources
Report footer	Keep weekly reports concise; often, a single page is enough. For space flight projects, the report seldom is this short, so include a page number in the footer Give point-of-contact information with the page number, so readers can more easily get clarification if the report's pages become separated from the title

The weekly report gives stakeholders a concise summary of activities

22.3.7.5 Meet Reporting Requirements for Technology Projects

New technologies drive innovation, which in turn drives how we will remain competitive in the modern world. Therefore, technologists have their own reporting requirements to gain the kinds of advantages presented in Table 22.10. Working together to ensure a coordinated approach to reporting, the technologists and project managers should discuss the requirements levied on each and factor these into the project's business rhythm. In certain instances, the reporting process will not be under their control, which calls even more strongly for coordination and cooperation.

Avoid the natural tension that reporting can cause. In general, reporting can cause tension between two parties, including project managers and technologists. First, as noted above, technologists may feel constrained by the relatively

inflexible reporting formats. Second, reporting for a space flight mission may be more frequent than for a technology project, so it is understandable for the technologist to think, 'Didn't we just do this?' Third, projects are seldom on schedule, and this is no less true for technology projects. These three conditions can easily produce frustration and tension. Add to them any team's natural bias towards optimistically assessing a project's present-day and near-term status and it is easy to see how reporting can be a challenge. Stakeholders have reason to doubt the space flight project if it always presents a rosy picture but then cannot deliver the products. It is far better to have candid discussions on the project's true state, including the technology effort, so that we can present the necessary risk-weighted assessments.

Report the 'right' message. The best project managers have learned early in their careers that reporting is both art

Table 22.9 Content of the monthly status report

Monthly status report section	Content
Title page	Include at least the following information:
	Project name and organizational code
	Review name (Monthly Status Review)
	Date
	Current project phase (Pre-A, A, B, C, D, E)
	Current planned launch date
	Names of key people including the
	Project Manager
	Deputy Project Manager
	Deputy for Resources
	Project Scientist or PI
	Systems Assurance Manager
	Lead Mission Systems Engineer
	Major contractors
	WBS or project numbers
Fever chart	Group topics as the project requires but usually include all ground elements, instruments, and flight systems
	Include a legend and summary assessment at bottom
	Be sure readers can distinguish between the red, yellow, green assessments on the black/white copies
Problems and issues	Have one problems/issues chart for each current non-green assessment on the fever chart
	Underline text changes the first month they are shown
	Keep the projected completion date for the issue's resolution current
	When a problem/issue is entirely resolved, say so in the current status section; shade or crosshatch the text in the upper part of the chart and show the color as green. Then, drop the chart from future packages
	The project manager decides whether to close a problem or issue. Do not ask to close it; simply say it has been closed and why. The stakeholders will let you know if they disagree
Risk matrix	Identify the project's top technical and programmatic risks (typically ten or fewer) in the 5×5 risk-matrix format
	Most projects track more risks than they show on the matrix, but stakeholders usually don't need to see more than about ten. If you're carrying more than ten in the red to yellow category (not to be confused with issues), we suggest identifying numbers 11 and beyond in the report's backup section
	At the review, focus on the reds and yellows and on changes from the previous month. Show green risks but do not discuss them unless an audience member asks you to
	Do not confuse risks with problems or issues from the previous section. Risks are the bad things that might happen. Problems or issues are things that have happened
	Make sure you show both the rank number and the risk ID
	If a risk has changed cells since the previous month, show an arrow going from where it had been to where it is now. You need not show what the rank number was in the previous month

(continued)

Table 22.9 (continued)

Monthly status report section	Content
Risk focus	<p>Provide this information only for risks falling into the red and yellow areas of the risk matrix</p> <p>Express the risk statement as an 'If/Then' or a 'Given that/There is a possibility that' statement</p> <p>Underline any wording changes from the previous month</p> <p>Typically, the project manager needs to speak only to the new risks and to changes since last month</p> <p>Make sure you include the rank number, risk ID, trend arrow, and risk criticality (high, medium, low) on this page</p> <p>If the approach to dealing with the risk is 'mitigate', the information in the status column will suffice to indicate the types of mitigations being done</p> <p>If the approach to dealing with the risk is 'watch', offer a very brief rationale in the approach column</p> <p>If your approach is to mitigate the risk, include the projected closure date in the 'rank' column</p>
Residual risks	<p>Provide this information only for missions within 6 months of launch</p> <p>Show where the residual risks (ones that won't be mitigated before flight) fall within the standard risk matrix</p> <p>Include a table defining each risk and the rationale for accepting the risk as is</p>
Significant progress	<p>Summarize in bullet form the major project accomplishments since the last report</p> <p>Be brief; keep the level of detail at an appropriate level. One or two charts should suffice</p>
Project scientist's or principal investigator's perspective	<p>Summarize the science team's recent and upcoming activities</p> <p>Present the project scientist's watch list, concerns, and issues</p> <p>Report performance status on key science requirements. In case of serious issues with performance, include another chart or two to explain</p>
Status of open actions	<p>Describe the status of any open actions from stakeholders at previous reviews</p> <p>Identify the source and number of every action item</p> <p>If you believe an action is ready for closure, recommend closure to the stakeholders (recognizing the group may not agree). Until you receive positive confirmation (oral or otherwise) from the meeting's chair, keep the action open</p> <p>If the all agree to close an action, don't show it at future MSRs</p>
Master schedule	<p>Show overall mission schedule through launch</p> <p>Indicate mission phases and system-level reviews (precursors to key decision points)</p> <p>Vary content from mission to mission, but typically include spacecraft, instruments, observatory, ground system, and launch vehicle</p> <p>Clearly show critical path</p> <p>Because this chart has a lot of detailed information, speak about something that has changed from the previous month</p>
Critical milestones	<p>Show the status of key milestones over a one-year period (3 months in the past, nine in the future)</p> <p>Circle milestones that have changed from the previous month</p> <p>Explain changes in milestones from the previous month by line number. Place explanations at the bottom of the page or on a separate page if you need more room</p>
Cumulative milestone	<p>Show the completion status of key milestones over about one to 2 years</p> <p>Include critical milestones, hardware deliveries, and items retaining significant risk</p> <p>Plot enough milestones to be able to show progress from month to month, but not so many you make it a bookkeeping nightmare</p> <p>These charts are particularly appropriate during the build-up and test phase. Do not use a milestone chart if the project is close to launch and waiting for a launch vehicle or so early in the life cycle that it doesn't make sense</p>

(continued)

Table 22.9 (continued)

Monthly status report section	Content
Cost and obligation status	Show separate charts for cost and obligation status
	For each cost element show plan, actual, delta, programmatic delta; explain variances
	Show overall total costs
Contingency status	Show total contingency, encumbrances, contingency through encumbrances, liens, contingency through liens by fiscal year, and contingency through threats
	Threats are ALL risks (red, yellow, and green) with a likelihood of 2 to 4 that have potential cost impacts
	Liens are ALL risks with a likelihood of 5 (have become issues) that have potential cost impacts
	Encumbrances are risks or issues that are fully realized and have an effect on final cost
	Only in unusual circumstances does a threat or lien not result from risk
	Calculate threats and keep books based on the project's assessment of likelihood consistent with placing them on the risk matrix according to the following formula: expected value of the threat = probability (P) x estimated cost impact
	P = 20 % for L of 2 (low)
	P = 40 % for L of 3 (moderate)
	P = 60 % for L of 4 (high)
	Lien for L of 5 (very high)
	Calculate liens and keep book at 100 % of the estimated cost impact
	Keep book on encumbrances at 100 % of the final cost impact
	Schedule slack and estimate at complete trend
Show schedule slack in working days, not calendar days	
For the slack trend, plot the slack itself not slack change. Where relevant, plot the 'one month per year' standard for comparison	
For the estimate at completion (EAC) trend, plot either the EAC itself or change in it. In either event, make sure the chart shows original and current EAC	
Open and closed review action status	Typically supplied to the project by the Independent Review Board Chair, this chart should reflect the agreed-on status of actions opened at past project reviews
	Tracks progress on closing out actions from system-level reviews, such as the PDR, CDR, PER, and PSR. Do not use it to track RFA closure from peer reviews or other less formal, lower-level reviews
	If the project has open actions that are both late in closing and 'critical', the project manager should address their status
Education and public outreach	Summarize ongoing or individual education and public outreach activities
	May include publications, press releases, conferences, displays, interviews, videos, school visits, education packages, and so on
	Include required performance metrics and statistics on EPO funding
	Typically, do not spend much time presenting this chart, unless it presents specific concerns or audience members ask about it

(continued)

Table 22.9 (continued)

Monthly status report section	Content
Project resources (cost and schedule reserves, power and mass margins)	<p>Explain deviation from standard values for budget reserves, schedule slack, mass margin, and power margin in the box at bottom of chart</p> <p>Standard values are as follows:</p> <p>Budget reserves: 20 %</p> <p>Schedule slack: 1 month/year before Observatory integration and test (I&T), 2 months/year during I&T, 1 week/month at launch site</p> <p>Mass and power margin:</p> <p>About 25 % at PDR</p> <p>About 20 % at CDR</p> <p>About 15 % at PER</p> <p>About 10 % at end of testing</p> <p>Adding the percentages of estimated, calculated, and measured mass is useful</p>
Executive summary quad chart'	<p>Typically used by senior management, who want only a summary of the report</p> <p>In 'quad chart' format include: mission objectives, mission partners and key vendors, instruments or technology being flown, overall status. For every project whose overall status is other than 'green', include a return-to-green recovery plan in the same quadrant</p>

Space flight projects must present detailed status reports, usually monthly. This report addresses all project aspects: programmatic, technical, budget, schedule, and risk

Table 22.10 Drivers for technology reporting

Potential reporting requirement	Context
Identify valuable technologies	Through active reporting of technology activities and progress, organizations and stakeholders can compile the most important activities to advocate. Advocacy can lead to continued funding support, diversify the user base and, if a government is involved, spin off the technology to the private sector
Assess commercial potential	One of the technologist's goals may be the widespread adoption of a new approach, capability, or function. Anticipate reporting to support a decision on when to assess whether or not to spin off a technology. These assessments are typically routine and therefore have predefined costs
Protect intellectual property	Account for cost of applying for domestic and international patents, copyrights, and related protection measures. When budget is constrained, the space flight project must fund the application, so include property protection in estimation templates
Meet export-control regulations	<p>Because technological advancements contribute strongly to a competitive advantage, they are often part of protection measures at some point in a project's life cycle. Anticipate reporting to support determining appropriate protection measures</p> <p>You won't be able to define costs for this type of reporting until the technology becomes part of a space flight mission. Then you can determine reporting costs</p> <p>Changes in partnership arrangements also may drive a corresponding change in reporting requirements. Include this possibility in the cost-benefit analysis you use to determine the efficacy of a change in the first place</p>
Report on organization's performance	You may have to file reports with the hosting organization so the organization can meet reporting requirements. In these instances, you will typically know the frequency and format, so you can incorporate the reporting cost in the technology project's budget estimates

Reporting requirements have many sources, and effective reporting can be helpful for any technology project

and science. The space flight project must give stakeholders a clear, consistent assessment of all the project's activities in each reporting cycle. Because people will read status reports at a future time and without the team's direct participation, the project's message can easily become insular and constrained. In short, the project reports what it would want to hear, not what the stakeholders want to receive.

Thus, many projects deliver only half the information that they should be sending out.

The most common error is failing to follow the formula: complete message = fact + consequence. Presenters often mention a project event but then fail to explain cogently why it was important. By not delivering the 'consequence' part of the message (positive and negative), these presenters

disregard the stakeholders' need to understand why the event mattered in the first place. In short, reporting means building advocacy over time, so that project managers and technologists can influence meetings and conversations they will never attend. This influence often can help a project succeed.

Nothing is worse for the reputation of project managers, technologists, or their projects than being questioned on the veracity of their reports. Managers prize being seen by the organization and stakeholders as an accurate source of reporting. At the start, everyone wants to do this, but reports diverge from reality for several reasons. Human nature is the main cause: the natural inclination to over-play the positive and under-play the negative. Understanding this begs the question, 'How should one proceed?' Among the best answers is that project manager's use reporting to 'set expectations' and balance stakeholders' expectations with their own. Then, they take the opportunity to explain the plans that are in place to keep the project on track, recover from a problem, or show why reevaluating the project's approach is not necessary.

22.3.8 Documenting Lessons Learned

Organizations, stakeholders, project managers, technologists, and their teams want to repeat past successes. Although their intent is clear, their ways of reaching this goal are not. In reality, repeating past successes can be elusive because no two projects are exactly the same. It does happen, but not with the desired frequency and certainty, or without stressing stakeholders to their limits. Clearly, no single recipe or formula helps us address this area.

As a result, organizations have defined processes to capture a team's experiences and packaged the information in a format that other teams can exploit. This practice is commonly referred to as capturing lessons learned. Although this process is useable at any time in the life cycle, organizations usually mandate it as part of a space flight project's closeout activities. Whenever it takes place, technologists should expect to be included, and project managers should encourage their participation.

22.3.8.1 Balance Lessons-Learned Databases with Interpersonal Forms

Although recording lessons learned is an excellent way to examine and capture actual events, it has inherent weaknesses. Thus, multiple processes are in place to help, ranging from lesson databases to real-time knowledge exchanges known as 'master's forums'. On the one hand, database solutions allow practitioners to search through the information on their own, when they need an insight and as

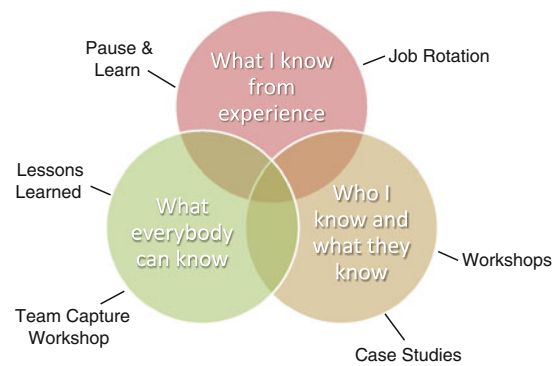


Fig. 22.11 Flight Projects Directorate's consolidated learning model. Building on the work of GSFC's Chief Knowledge Officer, the Flight Projects Directorate has audited its learning processes and intends to augment them as necessary in order to support learning for practitioners and organizations

often as they wish. However, the amount of data that an organization collects will grow until it eventually becomes difficult to sift through. Unless designers build the correct schema up front in order to control the data record format, keywords, and other data-management characteristics, practitioners will find the process cumbersome and inefficient as the database matures. On the other hand, master's forums are a great way to understand the context behind a lesson, the environment that drove the event, and any responses to it. However, subject matter experts are invariably very busy, so interactions with them are limited, and usually only a few practitioners can benefit from the exchange. The answer lies somewhere in the middle: a balanced approach between databases and inter-personal interactions.

22.3.8.2 Take Advantage of Consolidated Learning

GSFC's Flight Projects Directorate (FPD) has taken the lessons-learned concept a step further by committing themselves to being a 'learning organization'. This means they emphasize more than just conveying experiences between practitioners or teams by moving toward a grander vision that focuses on the concept of consolidated learning. This concept has arisen through extensive collaboration with Dr. Edward Rogers, GSFC's Chief Knowledge Officer. Organizations first must understand how practitioners learn. Although several models explain the learning process, Fig. 22.11 illustrates the model that GSFC uses for this collaborative effort.

Using the model in Fig. 22.11, the FPD's Advanced Concepts and Formulation Office audited GSFC's process for new mission competitions. The results showed that we need to add two new activities: 'pause and learn' and the 'kickoff workshop'. Because the concept of a kickoff

Fig. 22.12 Competitive new business process incorporating consolidated learning. To balance personal interactions and data mining, we must blend process and practitioner learning

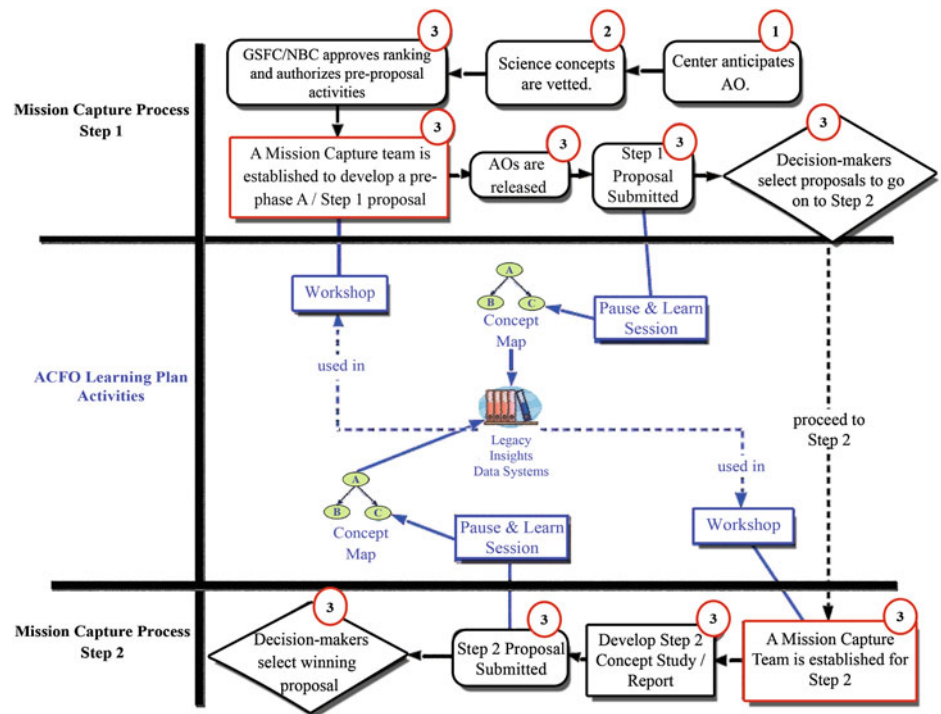
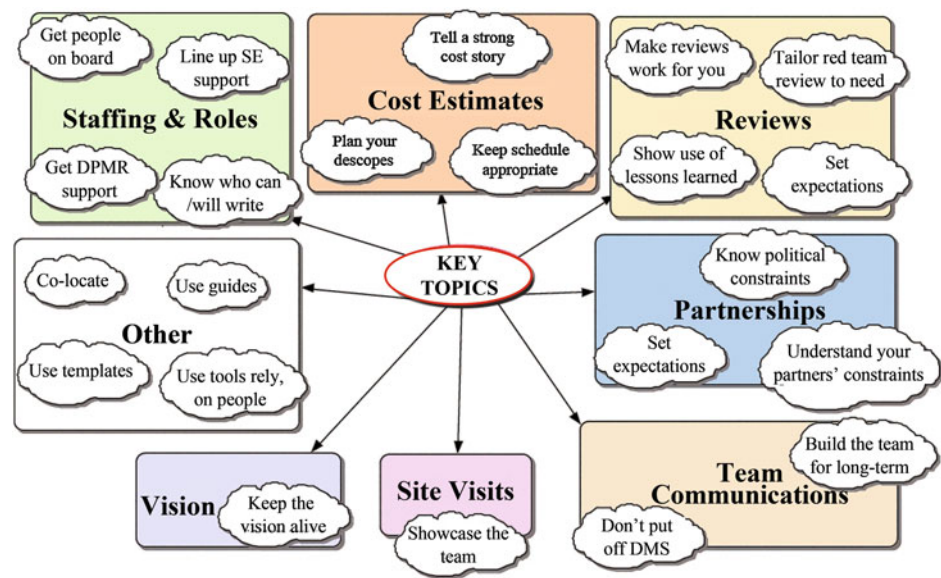


Fig. 22.13 Insight-based model for documenting practitioner learning. How best to aggregate the data challenges all data-collection activities. The Flight Projects Directorate uses an insight-based data scheme for new business



workshop is familiar to most people who have worked on business projects, we will discuss here only the pause and learn activity. Figure 22.12 illustrates the augmented competitive process that is now in the FPD.

The foundation for consolidated learning within the FPD is the pause and learn [17]—an interpersonal activity in which team members gather and focus on recently experienced events. Either the GSFC’s Chief Knowledge Officer or the FPD’s own knowledge manager leads this activity. A

pause and learn can occur at any time that practitioners or teams consider it appropriate to step back and discuss recent events. These ‘events’ may be a short-term activity such as preparing for an independent review, a longer activity such as a single life cycle phase, or the entire duration of a space flight mission.

The power in a pause and learn activity comes from the way that it directly affects participants in many ways. First, each person can ‘consolidate’ his or her learning over the

period being examined, which directly influences the ‘what I know’ aspect in Fig. 22.11. Second, other members of the team get a chance to correct the underlying assumptions for why something happened in the way that it did. As a result, they ensure that the individual practitioner learns the right lesson. This is a very important point. It is excellent to observe something directly, but if practitioners incorrectly understand the underlying cause, they will find it difficult to use this information later in order to anticipate a needed action. Interacting with the rest of the team enhances practitioners’ ability to respond correctly to future events. Third, pause and learn benefits the organization that learns from its practitioners’ experiences. This organizational knowledge can then flow into a database for use by other employees assigned to similar roles and responsibilities. Pause and learn is a powerful concept that we should consider to be part of any learning process.

Insights collected during a pause and learn activity represent valuable information for the organization and new practitioners. Therefore, we need databases to store, group, and serve up the information. How an organization handles this task depends on many factors, including project size for the database, diversity of the insights being uncovered, and the sensitivity of the information. The FPD has chosen to use an insight-based model for aggregating the information taken from its new business-focused pause and learn events. Figure 22.13 shows an example of this scheme. Managers then review the information with new practitioners during orientation and make it available for any future inquiries.

Whereas technology projects tend toward reduced documentation of requirements, costs, or schedule, space flight projects swing to the other extreme. So technologists should anticipate the call for lessons learned and capture them in a format that matches the project’s database. More importantly, technology projects should collect the data when its teams are doing the work, so that it will be of the highest quality but take the least time to produce. Although this approach may add a little up-front work on the team, it will pay significant dividends later, when the space flight project asks for it in order to do analyses and trade studies. In short, technologists should include data collection in their own project activities for the technology project’s own benefit. Not only will the data be readily available, it may even make it easier for the space flight project to defend its decision to use the technology in the first place!

The best project managers manage with a sense of urgency but without a sense of panic. These are different characteristics and must be understood. Project managers hold the project’s cost and schedule reserve and continually evaluate all risks facing it, not just those from the technology effort. Technologists should manage their element’s development with equal urgency.

22.4 Summary

A space flight mission involves many relationships. Of these, the relationship between project managers and technologists is one of the most critical. It can also be one of the most challenging because these two groups live in different worlds. Technologists live in the world of ‘innovation and firsts’, where intelligent people strive always to invent a newer and better way to carry out a function or capability. By contrast project managers are psychologically predisposed to ‘bound and deliver’—to deliver a product on time and on schedule by avoiding requirements creep, mitigating risks as early as possible, and holding costs to the original estimates. The adage ‘better is the enemy of good enough!’ best describes a successful space flight project manager’s mindset.

Both sides of this critical relationship must understand the other. Otherwise, technologists may reasonably conclude that space flight project managers are overly conservative at times, whereas the latter may conclude that technologists are never ready to stop innovating. Effective collaboration should produce a dynamic tension that can benefit the combined project by forcing both sides to work to a level of ‘pragmatic excellence’ that they might not otherwise achieve. Still, both parties must give the relationship genuine effort in order to ensure that it doesn’t skew too far to one side or the other.

The successful launch, checkout, and transition to normal operation of any spacecraft are exhilarating experiences. This is especially true when the technology and space flight projects are able to experience these milestones as a fully integrated, unified project team.

References

1. A Guide to the Project Management Body of Knowledge, Fourth Edition, Project Management Institute, PMI Publications, 2008
2. “Press Release: The 2006 Nobel Prize in Physics”. Nobelprize.org. 7 Sep 2012 http://www.nobelprize.org/nobel_prizes/physics/laureates/2006/press.html
3. United States. National Aeronautics and Space Administration. NASA Procedural Requirement (NPR) 7120.8, 2008
4. Chesley J, Larson W, McQuade M, MenradR, editors. Applied Project Management for Space Systems. New York: McGraw Hill; 2008
5. Cleland D, Ireland L. David Cleland’s Project Management—Strategic Design and Implementation. 4th ed. New York: McGraw Hill; 2002
6. Meredith J, Mantel S Jr, Jack R. Meredith’s Project Management—A Managerial Approach. 7th ed. New Jersey: John Wiley & Sons; 2009
7. United States. National Aeronautics and Space Administration. NASA Procedural Requirement (NPR) 7120.5e, 2012
8. United States. Department of Defense. Directive number 5000.1, May 2003

9. Burnett, David W., Sauve, Steve, "Project Management Maturity," *Applied Project Management for Space Systems*, Chesley, Larson, McQuade and Menrad [eds], 1st edn, McGraw Hill, 2008, pp. 649-672
10. Ibid [7]
11. Ibid [4]
12. A guide to the Project Management Body of Knowledge, Fourth Edition, Project Management Institute, PMI Publications, 2008
13. Ibid [4]
14. Casani, John R., Gavin, Thomas R., Sasaki, Chester N., "Mission Paradigms," *Applied Project Management for Space Systems*, Chesley, Larson, McQuade and Menrad [eds], 1st edn, McGraw Hill, 2008, pp. 682-698
15. United States. National Aeronautics and Space Administration. NASA Procedural Requirements (NPR) 8000.004A: Agency Management Management Procedural Requirements, December 2008
16. United States. National Aeronautics and Space Administration. NASA Procedural Requirements (NPR) 7120.4: Risk Classification for NASA Payloads, Appendix A, July 2008
17. "Knowledge Management" by Edward W. Rogers in "System Health Management with Aerospace Applications" by Stephen B. Johnson (ed). 2011 Wiley, page 69

Tanja Masson-Zwaan and Richard Crowther

The aim of this chapter is to provide some insight into the legal aspects of space activities from the perspective of a non-lawyer. Engineers and scientists increasingly express an interest in the legal aspects of space activities. Law is sometimes seen as barring technological progress by imposition of prescriptive rules and regulations, but the understanding that law can also protect and promote the interests of space science and technology is gaining ground.

Interaction between the fields of space science and technology and space law is important; a lawyer cannot make a good law if he has no idea about the technological aspects involved. Likewise, it is unwise to make scientific innovations without some knowledge about the regulatory framework that may embrace them. In other words, space is by definition a field of activity where interaction between different disciplines is desirable and in fact necessary, and the multidisciplinary content of this handbook bears witness to that.

In this context, this chapter provides an overview of some of the main principles of international and national space law. In addition, it addresses the legal aspects of two established activities, namely exploration and exploitation of space, and the emerging commercial human space flight industry. Other topics would certainly also merit attention, for instance the increasing problem of space debris, but it is impossible to address all of them in detail within this context.

T. Masson-Zwaan (✉)
University of Leiden, Leiden, Netherlands
e-mail: t.l.masson@law.leidenuniv.nl

R. Crowther
UK Space Agency, Swindon, UK

M. Macdonald and V. Badescu (eds.), *The International Handbook of Space Technology*, Springer Praxis Books, DOI: 10.1007/978-3-642-41101-4_23,
© Springer-Verlag Berlin Heidelberg 2014

23.1 International Space Law

Perhaps the first question after finding out that ‘space law’ exists is where outer space actually begins. This is an understandable question; however, despite the clear engineering definition given in [Chap. 2](#), there is no firm answer. The topic has been debated in the United Nations (UN) for several decades, but no agreement has been reached so far. Various approaches and many theories exist and this is not the place to address them all, except to say that with the advent of commercial human suborbital flights (which will be addressed further in this chapter), the time may soon come when a boundary between airspace and outer space must be defined. To date, there has been no such need, as activities were either very clearly a space activity, such as the building and operation of the ISS at an altitude of about 400 km, or they were very clearly aviation, such as a commercial flight between Glasgow and Amsterdam at an altitude of approximately 10 km.

The importance of this delimitation issue is demonstrated by the fact that there is a fundamental difference in the regimes governing air space and outer space. The first is subject to sovereignty of the underlying state, whereas in outer space a regime of ‘freedom’ exists (be it with certain limitations, of course), and no state is allowed to claim sovereignty over outer space or the celestial bodies.

The drafting of outer space law was initiated immediately after the launch of the first object into outer space, as states were from the start convinced that regulation of man’s activities in outer space was necessary in order to ensure that outer space would be used for peaceful purposes and in an orderly manner.

The basis for the space flight regulatory environment is derived from Treaties and Principles developed by the United Nations. Since 1961, issues relating to the use of outer space have been dealt with through the United Nations Committee on the Peaceful Uses of Outer Space

(UNCOPUOS). The Scientific and Technical Subcommittee of COPUOS addresses related technical issues, whereas the Legal Subcommittee of COPUOS deals with legal matters. The executive function of UNCOPUOS is supported by the United Nations Office for Outer Space Affairs (UNOOSA).

UNCOPUOS, established in 1958 first as an ad hoc and later as a permanent committee of the United Nations, initially had around twenty member states, which enabled the committee to reach consensus relatively easily. This resulted in the adoption of five UN Treaties between 1967 and 1979, which set the scene for the activities of man in outer space.¹

- The ‘Treaty on principles governing the activities of states in the exploration and use of outer space, including the Moon and other celestial bodies’ of 1967 (Outer Space Treaty), serves as the ‘Constitution’ of space law.
- The ‘Agreement on the Rescue of Astronauts, the Return of Astronauts, and the Return of Objects launched into Outer Space’ of 1968, deals mainly with the legal status of astronauts in case of an accident (Rescue Agreement).
- The ‘Convention on international liability for damage caused by space objects’ of 1972, addresses the question of liability in case of damage caused by a space object (Liability Convention).
- The ‘Convention on Registration of Objects launched into Outer Space’ of 1976, creates an obligation to register objects launched into space both with the UN and at the national level (Registration Convention).
- The ‘Agreement Governing the Activities of States on the Moon and Other Celestial Bodies’ of 1979, addresses the legal status of celestial bodies and specifically the exploration and exploitation of natural resources of celestial bodies (Moon Agreement).

The first three treaties were ratified by close to ninety states (100 for the Outer Space Treaty), the fourth by around fifty, and the last by only thirteen states so far. Major space powers such as the USA, Russia, China, India, Japan, France, the UK, Canada, and Germany have all ratified the first four treaties. None of these states has ratified the 1979 Moon Agreement. Several international intergovernmental organizations (such as ESA, EUMETSAT and EUTELSAT) have declared their acceptance of the rights and obligations under the treaties (this is possible for all but the Outer Space Treaty). Many countries have reflected their obligations under the treaties through the enactment of national legislation.

The membership of UNCOPUOS has now grown to some seventy states, including many more space ‘haves’ but also numerous space ‘have-nots’. Developing countries began to impose their view that outer space should be the

“Common Heritage of Mankind,” rather than the “Province of all Mankind”. This and other matters related to the Moon Agreement led to the move to adopt Principles rather than Treaties since the 1980s.

The relevant aspects of the four main international treaties in relation to space regulation are presented below. The Moon Agreement will be addressed in more detail in the next section on exploration and exploitation.

23.1.1 The 1967 Outer Space Treaty

The basic provisions contained in the Outer Space Treaty, although they were drafted more than 40 years ago in a field that was subject to fast and profound technological advances, are still relevant today and are broad enough to address a wide range of space activities—even though admittedly it has become necessary to draft additional rules and regulations to elaborate on the principles contained in the Outer Space Treaty. The first and possibly the most important principle of space law is contained in Article I of the Outer Space Treaty. It reads as follows

The exploration and use of outer space, including the Moon and other celestial bodies, shall be carried out for the benefit and in the interests of all countries, irrespective of their degree of economic or scientific development, and shall be the province of all mankind. Outer space, including the Moon and other celestial bodies, shall be free for exploration and use by all States without discrimination of any kind, on a basis of equality and in accordance with international law, and there shall be free access to all areas of celestial bodies. There shall be freedom of scientific investigation in outer space, including the Moon and other celestial bodies, and States shall facilitate and encourage international co-operation in such investigation.

Of course, the concepts are not clearly defined and can be subject to varying interpretations—but the general idea is clear: use of space should somehow benefit humanity.

The second-most important principle of space law is contained in Article II, which declares that outer space and celestial bodies cannot be subject to appropriation by any means. It reads as follows:

Outer space, including the Moon and other celestial bodies, is not subject to national appropriation by claim of sovereignty, by means of use or occupation, or by any other means.

This means that there is no ‘territorial jurisdiction’ in outer space, unlike on Earth or in the airspace above the territory of a state. Thus, the planting of a US flag on the Moon during Apollo-11 in 1969 did not imply that the Moon had become US territory. It is forbidden to claim ownership of any part of outer space, and this applies not only to states but also to private entities, contrary to what is sometimes argued, because there is no sovereign authority that has competence to confer titles of ownership.

¹ All texts, official titles and sources can be consulted on the useful website of the Office for Outer Space Affairs in Vienna, the UN office supporting the work of COPUOS. See <http://www.oosa.unvienna.org>, especially under ‘Space law’.

Another important feature is that activities must be carried out in accordance with international law, including the UN Charter, in the interest of maintaining international peace and security and promoting international cooperation and understanding (Article III of the Outer Space Treaty). This means that provisions of the UN Charter such as Article 2.4 and Article 51 on the duty to refrain from the threat or use of force and the inherent right of self-defense are equally applicable to man's activities in outer space. Article IV further addresses the military uses of space, and includes a prohibition on placing nuclear weapons and weapons of mass destruction anywhere in outer space; it seems however that military use is not absolutely prohibited, as the requirement to use space "exclusively for peaceful purposes" only applies to the Moon and other celestial bodies, and not to outer space *per se*.

Article V then addresses astronauts, and declares that states should regard them as "envoys of mankind". The Treaty does not provide any guidance on the meaning or implications of this term, except to say that both states parties and astronauts of other states should render all possible assistance to astronauts in distress.

The treaties also contain rules concerning responsibility and liability (Article VI and VII of the Outer Space Treaty, further elaborated in the Liability Convention). A state is responsible for "national activities" in space, and a launching state is liable for damage caused by its space object to another state or its natural or juridical persons, whether that damage occurs in space, in the air or on the ground. Article VI reads

States Parties to the Treaty shall bear international responsibility for national activities in outer space, including the Moon and other celestial bodies, whether such activities are carried on by governmental agencies or by non-governmental entities, and for assuring that national activities are carried out in conformity with the provisions set forth in the Treaty. The activities of non-governmental entities in outer space, including the Moon and other celestial bodies, shall require authorization and continuing supervision by the appropriate State Party to the Treaty. When activities are carried on in outer space, including the Moon and other celestial bodies, by an international organization, responsibility for compliance with this Treaty shall be borne both by the international organization and by the States Parties to the Treaty participating in such organization.

And Article VII states

Each State Party to the Treaty that launches or procures the launching of an object into outer space, including the Moon and other celestial bodies, and each State Party from whose territory or facility an object is launched, is internationally liable for damage to another State Party to the Treaty or to its natural or juridical persons by such object or its component parts on the Earth, in air or in outer space, including the Moon and other celestial bodies.

It must be noted that space law only has a system of state liability, i.e. a private entity or a natural person cannot present

a claim based on the Treaty against another state directly under the Treaties but must be represented by its state; this is yet again an important difference with the system of air law, where for instance a passenger who suffered damage on board an aircraft can present a claim for damage directly to the operator of the aircraft. Of course a space passenger could still present a claim against a space operator under national law (breach of contract, tort), for instance in the jurisdiction of the space object or another appropriate jurisdiction.

A system of dual registration has been elaborated, whereby states register an object that they launch into outer space both in a national register and in a central UN register, and jurisdiction and control are exercised by the state of registry (Article VIII of the Outer Space Treaty, further elaborated in the Registration Convention). The requirements for registration are, however, not very detailed and there have recently been discussions about the need to improve them. The reason is that identification of defunct objects or parts of such objects could be easier if the details given during registration were more elaborate. This would be helpful in dealing with the growing problem of 'space debris,' although identification is only one of the problems related to space debris, and pieces that reenter the Earth's atmosphere are usually large enough to identify (another major problem posed by space debris being the establishment of fault for damage caused in outer space itself). Regarding registration, Article VIII of the Outer Space Treaty declares that

A State Party to the Treaty on whose registry an object launched into outer space is carried shall retain jurisdiction and control over such object, and over any personnel; thereof, while in outer space or on a celestial body. Ownership of objects launched into outer space, including objects landed or constructed on a celestial body, and of their component parts, is not affected by their presence in outer space, or on a celestial body, or by their return to the Earth. Such objects or component parts found beyond the limits of the State Party to the Treaty on whose registry they are carried shall be returned to that State Party, which shall, upon request, furnish identifying data prior to their return.

Furthermore, it is important to mention Article IX in view of the next section on exploration and exploitation. Article IX is the only article that addresses the problem of contamination. It states that

In the exploration and use of outer space, including the Moon and other celestial bodies, States Parties to the Treaty shall be guided by the principle of co-operation and mutual assistance, and shall conduct all their activities in outer space, including the Moon and other celestial bodies, with due regard to the corresponding interests of all other States Parties to the Treaty. States Parties to the Treaty shall pursue studies of outer space, including the Moon and other celestial bodies, and conduct exploration of them so as to avoid their harmful contamination, and also adverse changes in the environment of the Earth resulting from the introduction of extra-terrestrial matter, and where necessary, shall adopt appropriate measures for this purpose. If a State Party to the Treaty has reason to believe that

an activity or experiment planned by it or its nationals in outer space, including the Moon and other celestial bodies would cause potentially harmful interference with activities of other States Parties in the peaceful exploration, and use of outer space, including the Moon and other celestial bodies, it shall undertake appropriate international consultations before proceeding with any such activity or experiment. A State Party to the Treaty, which has reason to believe that an activity or experiment planned by another State Party in outer space, including the Moon and other celestial bodies would cause potentially harmful interference with activities in the peaceful exploration and use of outer space, including the Moon and other celestial bodies, may request consultation concerning the activity or experiment.

And lastly, Article XI must be mentioned, which declares that

In order to promote international co-operation in the peaceful exploration and use of outer space, States Parties to the Treaty conducting activities in outer space, including the Moon and other celestial bodies, agree to inform the Secretary-General of the United Nations as well as the public and the international scientific community, to the greatest extent feasible and practicable, of the nature, conduct, locations, and results of such activities. On receiving the said information, the Secretary-General of the United Nations should be prepared to disseminate it immediately and effectively.

In summary, the major points addressed by the Treaty are

- The exploration and use of outer space shall be carried out for the benefit and in the interests of all countries and shall be the province of all mankind.
- Outer space shall be free for exploration and use by all States.
- Outer space is not subject to national appropriation by claim of sovereignty, by means of use or occupation, or by any other means.
- States shall not place nuclear weapons, or other weapons of mass destruction in orbit, or on celestial bodies, or station them in outer space in any other manner.
- The Moon and other celestial bodies shall be used exclusively for peaceful purposes.
- States shall be responsible for national space activities whether carried out by governmental or non-governmental activities.
- States shall be liable for damage caused by their space objects.
- States shall avoid harmful contamination of space and celestial bodies.
- States shall regard astronauts as envoys of mankind in outer space, and shall render to them all possible assistance in the event of accident, distress, or emergency landing on the territory of a foreign State, or on the high seas.

These issues are further developed by the subsequent treaties.

23.1.2 The 1968 Rescue Agreement

The Rescue Agreement elaborates on the provisions in Article V of the Outer Space Treaty. It is slightly more specific than Article V, but does not solve all the problems of interpretation. Interestingly, it does not use the term “envoy of mankind” as used in the Outer Space Treaty. On the other hand, it does not differentiate between professional astronauts or other passengers on board space objects, nor does it create special status or powers for the commander of a space object.

The pertinent elements of the Rescue Agreement are as follows. Article 5 of the Rescue Agreement declares that

- Each Contracting Party which receives information or discovers that a space object or its component parts has returned to Earth in territory under its jurisdiction or on the high seas or in any other place not under the jurisdiction of any State, shall notify the launching authority and the Secretary-General of the United Nations.
- Each Contracting Party having jurisdiction over the territory on which a space object or its component parts has been discovered shall, upon the request of the launching authority, and with assistance from that authority if requested, take such steps as it finds practicable to recover the object or component parts.
- Upon request of the launching authority, objects launched into outer space or their component parts found beyond the territorial limits of the launching authority shall be returned to or held at the disposal of representatives of the launching authority, which shall, upon request, furnish identifying data prior to their return.
- Notwithstanding paragraphs 2 and 3 of this article, a Contracting Party which has reason to believe that a space object or its component parts discovered in territory under its jurisdiction, or recovered by it elsewhere, is of a hazardous or deleterious nature may so notify the launching authority, which shall immediately take effective steps, under the direction and control of the said Contracting Party, to eliminate possible danger of harm.
- Expenses incurred in fulfilling obligations to recover and return a space object or its component parts under paragraphs 2 and 3 of this article shall be borne by the launching authority.

Article 6 of the Rescue Agreement states that

For the purposes of this Agreement, the term ‘launching authority’ shall refer to the State responsible for launching, or, where an international intergovernmental organization is responsible for launching, that organization, provided that organization declares its acceptance of the rights and obligations provided for in this Agreement, and a majority of the States members of that organization are Contracting Parties to this Agreement and to the Treaty on Principles governing the

activities of States in the Exploration and Use of Outer Space, including the Moon and other celestial bodies.

The Rescue Agreement is important in establishing international responsibilities in relation to the activities and property of other state actors in case of accidents, distress, emergency or unintended landings. It also addresses the legal status of the launching state, an important aspect of space regulation, and a role, which is further elaborated in Article I(c) of the Liability and Article I(a) of the Registration Convention.

In summary, the Rescue Agreement elaborates on those elements of the Outer Space Treaty that deal with assistance to astronauts and return of objects that crash or land in a foreign country or on the High Seas. The agreement provides that states shall take all possible steps to Rescue and assist astronauts in distress and promptly return them to the launching state. It also provides that states shall, upon request, provide assistance to launching states in recovering space objects that return to Earth outside the territory of the launching state.

23.1.3 The 1972 Liability Convention

The Liability Convention expands on Article VII of the Outer Space Treaty. The Convention has never been invoked in a court case, and hence its provisions, some of which are rather vague, have never had the benefit of being interpreted or clarified by case law. Some accidents could have led to claims under the convention, for instance, part of the cost incurred for cleaning up nuclear waste caused by the 1978 crash of Kosmos 954 on Canadian territory was reimbursed by the then USSR, but this was not done under the terms of the Convention (the USSR did not admit liability). The more recent collision between Iridium 33 and Kosmos 2251 in 2009 also did not lead to any liability claim under the Convention.

The Convention has a victim-oriented approach, and identifies several states as potentially liable 'launching states'. However, a problem lies in the fact that only states may present a claim. Private individuals or companies have no direct cause of action under the convention, but depend on their country to present a claim to (one of) the launching state(s).

In case a claim would be presented, the Convention provides detailed rules for the establishment of a claims commission (in fact, this is the longest of all five treaties, with 28 articles, as opposed to only ten for the Rescue Agreement).

The more pertinent elements of the Liability Convention are presented subsequently. In Article I, a number of important definitions are introduced

- The term 'damage' means loss of life, personal injury or other impairment of health; or loss of or damage to property of states or of persons, natural or juridical, or property of international intergovernmental organizations (it is unclear whether only direct damage is covered or also indirect damage, such as loss of revenue or emotional damage).
- The term 'launching' includes attempted launching.
- The term 'launching state' means a state which launches or procures the launching of a space object, or a state from whose territory or facility a space object is launched (it is rather unclear what 'procures' means, and states tend to attach different interpretations to this term).
- The term 'space object' includes component parts of a space object as well as its launch vehicle and parts thereof (here, it is not clear whether space debris still qualifies as a 'space object' to which liability attaches, even if the launching state can no longer exercise any control over it).

Article II states that

A launching State shall be absolutely liable to pay compensation for damage caused by its space object on the surface of the earth or to aircraft flight (this means that the launching state is liable irrespective of whether it was at fault; the rationale is that persons or property on earth or in the air should not be victimized by activities in outer space with which they have nothing to do, over which they have no control and little or no information).

And Article III declares

In the event of damage being caused elsewhere than on the surface of the earth to a space object of one launching State or to persons or property on board such a space object by a space object of another launching State, the latter shall be liable only if the damage is due to its fault or the fault of persons for whom it is responsible (in this situation, both states have engaged in space activity and must therefore assume the risks of an accident, except if one of them has committed a fault; obviously, in many cases it will be difficult or impossible to establish fault for an accident occurring in outer space).

Article IV states that:

In the event of damage being caused elsewhere than on the surface of the earth to a space object of one launching State or to persons or property on board such a space object by a space object of another launching State, and of damage thereby being caused to a third State or to its natural or juridical persons, the first two States shall be jointly and severally liable to the third State, to the extent indicated by the following:

- If the damage has been caused to the third State on the surface of the earth or to aircraft in flight, their liability to the third State shall be absolute;
- If the damage has been caused to a space object of the third State or to persons or property on board that space object elsewhere than on the surface of the earth, their liability to the third State shall be based on the fault of either of the first two States or on the fault of persons for whom either is responsible.

In all cases of joint and several liability referred to in paragraph 1 of this article, the burden of compensation for the damage shall be apportioned between the first two States in accordance with the extent to which they were at fault; if the extent of the fault of each of these States cannot be established, the burden of compensation shall be apportioned equally between them. Such apportionment shall be without prejudice to the right of the third State to seek the entire compensation due under this Convention from any or all of the launching States, which are jointly and severally liable.

Article V addresses joint liability

Whenever two or more States jointly launch a space object, they shall be jointly and severally liable for any damage caused.

A launching State, which has paid compensation for damage shall have the right to present a claim for indemnification to other participants in the joint launching. The participants in a joint launching may conclude agreements regarding the apportioning among themselves of the financial obligation in respect of which they are jointly and severally liable. Such agreements shall be without prejudice to the right of a State sustaining damage to seek the entire compensation due under this Convention from any or all of the launching States, which are jointly and severally liable.

A State from whose territory or facility a space object is launched shall be regarded as a participant in a joint launching.

Article VII addresses some exemption issues by declaring that

The provisions of this Convention shall not apply to damage caused by a space object of a launching State to:

- Nationals of that launching State;
- Foreign nationals during such time as they are participating in the operation of that space object from the time of its launching or at any stage thereafter until its descent, or during such time as they are in the immediate vicinity of a planned launching or recovery area as the result of an invitation by that launching State.

In summary, the Liability Convention introduces definitions for damage due to space activities and determines those 'launching States' that could be potentially liable for such damage and the mechanisms for such claims. Elaborating on Article VII of the Outer Space Treaty, the Liability Convention provides that a launching state shall be absolutely liable to pay compensation for damage caused by its space objects on the surface of the Earth or to aircraft, and liable for damage due to its faults in space. The Convention also provides for procedures for the settlement of claims for damages.

23.1.4 The 1975 Registration Convention

The last of the major outer space treaties is the Registration Convention. It addresses the important issue of notification

of activities to third parties, and establishes a key role for the United Nations through its Office for Outer Space Affairs (UNOOSA).

The Registration Convention provides that a launching state should furnish to the United Nations, as soon as practicable, the following information concerning each space object

- Name of launching state.
- An appropriate designator of the space object or its registration number; Date and territory or location of launch.
- Basic orbital parameters, including
 - Nodal period (the time between two successive north-bound crossings of the equator, usually in minutes).
 - Inclination (a polar orbit is 90° and equatorial orbit is 0°).
- Apogee (highest altitude above the Earth's surface).
- Perigee (lowest altitude above the Earth's surface).
- General function of the space object.

This information, although useful for identifying the launch of a space object, has limited operational, as opposed to the Two-Line Elements introduced in Sect. 2.2.3, value in determining the position of the space object once the initial injection into orbit has been performed. Although the adoption of UN General Assembly Resolution 62/101 of 17 December 2007 ('Recommendations on Enhancing the Practice of States and International Intergovernmental Organizations in Registering Space Objects') may increase the efficiency of the registration process, much uncertainty remains, for instance in the context of transfer of ownership of space objects.

23.1.5 Other Legal Instruments

Besides the treaties discussed above and the 1979 Moon Agreement, a number of other important principles produced by the United Nations have relevance to the space flight regulatory environment.

- Declaration of Legal Principles Governing the Activities of States in the Exploration and Use of Outer Space, adopted on 13 December 1963.
- Principles Governing the Use by States of Artificial Earth Satellites for International Direct Television Broadcasting, adopted on 10 December 1982.
- Principles Relating to Remote Sensing of the Earth from Outer Space, adopted on 3 December 1986.
- Principles Relevant to the Use of Nuclear Power Sources in Outer Space, adopted on 14 December 1992.
- Declaration on International Cooperation in the Exploration and Use of Outer Space for the Benefit and in the Interest of all States, Taking into Particular Account the Needs of Developing Countries, 1996.

These Resolutions do not have the same binding force as a Treaty, but since most of them were adopted by unanimity and have given rise to consistent state practice, at least some of the principles contained therein have become binding at international law as ‘international custom’.

Furthermore, the International Telecommunication Union (ITU) addresses the regulation of spectrum/orbit usage by spacecraft through a legal regime represented by the ITU Constitution, Convention and associated Radio Regulations. These instruments reflect the main ITU principles and lay down the specific regulations governing frequency spectrum allocations to different categories of radio communication services, rights and obligations of Member administrations to obtain access to the spectrum/orbit resources, and international recognition of these rights by recording frequency assignments, and as appropriate, orbital positions.

23.2 National Space Law

As mentioned earlier, a number of countries have reflected their obligations under the Outer Space Treaties through the enactment of national legislation. The focus in this chapter is mainly on general national laws setting out licensing procedures for private entities engaging in space activities. States may also have other national laws to regulate specific space activities, such as remote sensing or satellite telecommunications.

Norway introduced its ‘Act on launching objects from Norwegian territory into outer space’ as early as 13 June 1969. This was followed by Sweden in 1982 with its brief Act/Decree on space activities and associated licensing regime. The USA introduced its Commercial Space Launch Act (CSLA) in 1984, and the amended Commercial Space Launch Amendment Act (CSLAA) followed in 2004. The United Kingdom brought into force its ‘Outer Space Act’ in 1986. South Africa introduced its ‘Space Affairs Act’ in 1993, and Argentina developed its National Decree relating to the ‘Establishment of the National Registry of objects launched into outer space’ in 1995. The Russian Federation enacted its Decree and statute on licensing space operations in 1996, closely followed by the Ukraine with its ‘Ordinance of the Supreme Soviet of Ukraine on Space Activity Law’ again in 1996. Australia introduced its ‘Space Activities Act’ in 1998 and Brazil developed the ‘Brazilian Space Agency Administrative Edict’ in 2001. The Belgian ‘Law on the Activities of Launching, Flight Operations or Guidance of Space Objects’ followed in 2005, and the Netherlands ‘Space Activities Act’ was enacted in 2006 and came into force on 1 January 2008. France enacted the ‘Law on Space Operations’ in 2008. In December 2011, the Austrian ‘Federal Law on the Authorisation of Space

Activities and the Establishment of a National Registry’ was adopted.

States that do not have a national space law must still authorize and supervise their space activities by private entities, but do so on a case-by-case basis.

In the framework of this chapter, it is not possible to discuss all of these national space legislations, but a summary overview of the US and some European legislations is given below.

23.2.1 The United States

Within the USA, the regulation of space activities is governed separately for government and for commercial missions. The former are addressed as requirements by NASA, the Department of Defense and other non-regulatory organizations. Commercial missions are handled by the Department of Transportation, the Department of Commerce, and the Federal Communications Commission (FCC). Space regulations of the Federal Aviation Administration (FAA) of the Department of Transportation apply to launch vehicles and reusable spacecraft. FCC regulations apply to satellites licensed by the FCC on behalf of the United States, and to communications by non-U.S. licensed satellites with U.S. Earth stations. Commercial satellites performing remote sensing functions are licensed by the National Oceanic and Atmospheric Administration (NOAA) within the Department of Commerce.

23.2.1.1 Launches

The Commercial Space Launch Act (CSLA) is administered by the Office of Commercial Space Transportation, which is part of the FAA. The purposes of the CSLA are to safely open access to space and encourage private sector development, to simplify and expedite issuance and transfer of launch and reentry licenses, to promote safety, both of the public and of private property, and to strengthen and expand space transportation infrastructure. A license is required for launch and reentry in the United States, for launch and reentry by United States citizens outside of the United States, for launch and reentry by United States citizens outside the United States and outside territory of a foreign country unless the foreign country’s government has an agreement with the United States on jurisdiction over the launch or operation in question. A license is also required for launch or reentry by a United States citizen in a foreign country if the United States has jurisdiction by agreement with the government of a foreign country with respect to that launch.

The license application that is submitted to the FAA by the applicant is subject to a policy review, a safety review, and also a review of the environmental impact of the launch

activity or reentry activity. There is a requirement with respect to orbital debris mitigation, and also requirements relative to flight crew qualifications, training, and safety. The licensee must obtain third-party liability insurance or demonstrate financial ability to pay maximum probable loss arising from third-party claims. The maximum probable loss is established for each license by the FAA.

A launch-specific license authorizes a licensee to conduct one or more launches, having the same launch parameters, of one type of launch vehicle from one launch site. The license identifies, by name or mission, each launch authorized under the license. A licensee's authorization to launch terminates upon completion of all launches authorized by the license, or at the expiration date stated in the license, whichever occurs first.

A launch operator license authorizes a licensee to conduct launches from one launch site, within a range of launch parameters, and/or launch vehicles from the same family of vehicles transporting specified classes of payloads. A launch operator license remains in effect for five years from the date of issuance.

23.2.1.2 Remote Sensing

The Land Remote Sensing Policy Act of 1992 is administered by the National Oceanic and Atmospheric Administration of the Department of Commerce. Its purposes include the stimulation of commercial marketing for unenhanced remotely sensed data, the furthering of the long-term role of commercialization of land remote sensing, and the promotion of international trade and access to unenhanced data on a non-discriminatory basis. The Land Remote Sensing Policy Act also includes licensing and oversight responsibilities that are implemented by NOAA. A license is required to operate a private remote sensing satellite system and when the applicant makes the application, the applicant must provide NOAA with the orbit and data collection characteristics and any deviations therefrom. Included in the application also must be proper post-mission disposal. And, finally, Government approval is required for any significant or substantial agreement with a foreign entity.

23.2.1.3 Communications

The Communications Act of 1934 is administered by the Federal Communications Commission and it includes licensing and operating requirements for satellites and associated ground stations. The purposes of these provisions are to avoid radio frequency interference, to coordinate commercial satellite operations in the United States and to coordinate international satellite operations and use of the frequency spectrum along with the National Telecommunications and Information Administration and other United States agencies, through the International Telecommunication Union in

accordance with the ITU Radio Regulations. The Communications Act also includes orbital debris mitigation requirements, the purposes of which are to preserve continued affordable access to space, to continue the provision of reliable United States space-based services, and to help ensure the continued safety of persons and property in space and on Earth. To that end, an applicant is required to submit a mitigation plan along with its license application and this must include end of life operation requirements consistent with ITU requirements and ensure the discharge of on-board energy sources.

23.2.2 United Kingdom

Many of the processes and criteria used by the UK's Outer Space Act are similar to those used by the USA.

The Outer Space Act 1986 (OSA) is the legal basis for the regulation of activities in outer space (including the launch and operation of space objects) carried out by persons connected with the United Kingdom. The Act confers licensing and other powers on the Secretary of State acting through the UK Space Agency. The Act ensures compliance with UK obligations under the international conventions covering the use of outer space.

Under the legislation of the OSA, the Secretary of State shall not grant a license unless satisfied that the activities authorized by the license will not jeopardize public health or the safety of persons or property, will be consistent with the international obligations of the United Kingdom, and will not impair the national security of the United Kingdom.

Further, the Secretary of State requires the licensee to conduct operations in such a manner as to prevent the contamination of outer space or adverse changes in the environment of the Earth, and to avoid interference with activities of others in the peaceful exploration and use of outer space. For example, the Secretary of State may make regulations that

- Prescribe the form and contents of applications for licenses and other documents to be filed in connection with applications.
- Regulate the procedure to be followed in connection with applications.
- Authorize the rectification of procedural irregularities.
- Prescribe time limits for doing anything required to be done in connection with the application and providing for the extension of any period so prescribed.
- Require the payment to the Secretary of State of such fees as may be prescribed.

A license describes the activities authorized by it and shall be granted for such period, and is granted subject to such conditions, as the Secretary of State thinks fit. Further, a license may contain conditions that permit inspection by the

Secretary of State of the licensee's facilities and inspection and testing of the licensee's equipment. It also requires the licensee to provide such information as the Secretary of State thinks fit concerning the nature, conduct, location, and results of the licensee's activities.

The Secretary of State requires the licensee to insure against liability incurred in respect of damage or loss suffered by third parties, in the United Kingdom or elsewhere, as a result of the activities authorized by the license. Further, the licensee shall indemnify the government in the United Kingdom against any claims brought against the government in respect of damage or loss arising out of activities carried to which this Act applies. The requirement for insurance is prescribed in the Act, however determination of the level of insurance is at the discretion of the Secretary of State and during 2011 it was decided to reduce the level of insurance required to 60 million Euros for a standard launch/payload, consistent with many other regulatory authorities/launch service/insurance providers. In addition, the unlimited liability that the UK currently passes on to the license applicant should in the future be capped to the level of the insurance required. Such a change to the legislation of the UK OSA requires a Legislative Reform Order with Parliamentary oversight to come into effect.

The OSA provides the necessary regulatory oversight to

- Consider public health and safety, and the safety of property.
- Evaluate the environmental impact of proposed activities.
- Assess the implications for national security and foreign policy interests.
- Determine financial responsibilities and international obligations.

When a license is used with pre-set conditions, compliance monitoring is performed to ensure that a licensee complies with the Act, the regulations, and the terms and conditions set forth in its license. A launch licensee shall allow access by, and co-operate with, employees or other individuals authorized by the relevant agency to observe the activities of the licensee, or of the licensee's contractors or subcontractors, associated with the conduct of a licensed launch.

23.2.3 France

The Law on Space Operations of 2008 (Law No. 2008-518, of 3 June 2008, Relative to Space Operations) was enacted because France bears significant international responsibility as a launching state, particularly after the French state received a foreign operator's request to use the Kourou launch site in French Guyana. Accepting that request from a private operator meant assuming the eventual liability to third parties deriving out of launches carried out from French territory. Therefore, the need arose to regulate

authorization and supervision, as also the consequences of international liability, of space activities carried out by non-governmental entities in France.

The authorization system set up by the law is intended to allow the French Government to exercise control over the activities of operators likely to result in its liability as the launching state. With respect to the requirements that are necessary for an authorization by the competent authority, the law sets these main conditions

- Consistency with government policy, as well as financial and professional guarantees.
- Compliance of the envisaged activities with the Technical Regulations set down by the French Space Agency, CNES.
- Respect for the interests of national defense, as well as France's compliance with its international commitments.

The French national law requires insurance or another financial guarantee by the operator. Details of these financial guarantees are given in an Implementing Decree of 2009. According to Article 6, every operator who is subject to an authorization in application of this law must hold and keep in force, for the duration of the operation, an insurance policy or other kind of financial guarantee. The current insurance requirement is 60 million Euros, which is also the limit of the liability of the operator (this means that the state will cover any damage above this amount).

Compliance with Technical Regulations is mandatory for space operations by French space operators and for space operations from French territory. The space safety requirements and regulations governing procedures are based on national and international best practices and experience. A critical design review of the space system and procedures is to be carried out by the applicant, in order to verify compliance with the Technical Regulations. An independent technical assessment of the operation is delegated to CNES. The Technical Regulations are divided into three sections covering common requirements for the launch, control and return of a space object. A dedicated section will cover specific rules to be applied at the Guyana Space Centre in Kourou. The main topics addressed by the Technical Regulations are the operator safety management system; the study of risks to people, property, public health and the Earth's environment; the impact on the outer space environment, such as space debris generated by the operation; and planetary protection (Lazare).

23.2.4 Belgium

The Belgian space law provides for a flexible regime, since it has very generic terms for authorization (Law on the Activities of Launching, Flight Operations or Guidance of Space Objects, F. 2005-3027, September 2005). Article 5 establishes the general obligation to ensure safety of people

and property, the environment, the optimal use of airspace and outer space, the strategic, economic and financial interests of the Belgian state, and compliance with the international obligations of the Belgian state. Application of the law to specific cases is left to the Belgian authorities, which may attach further conditions to each particular authorization. There is no mandatory insurance under the Belgian law, and by a royal decree of 2008 the operator's liability is limited to 10 % of average operational turnover.

23.2.5 The Netherlands

The Netherlands Space Activities Act, in force since 1 January 2008, also establishes a flexible licensing system for private space operators, including all necessary requirements such as insurance and regulation of liability issues (Law on Rules Concerning Space Activities and the Establishment of a Registry of Space Objects, 2006). The Act contains a series of conditions to be complied with by operators, in relation to the safety of persons and property, environmental protection, public order and security, and financial security, as well as the compliance with the international obligations of the state. As usual in many national space laws, sufficient insurance coverage is a key requirement for granting a license. The amount of the required insurance is what the Minister considers to be the maximum possible cover for the liability arising from the space activities for which a license is requested, taking into account what can reasonably be covered by insurance. The liability of the license holder is limited to sum insured. The law does not apply to activities of Dutch citizens abroad.

23.2.6 Austria

The Federal Law on the Authorisation of Space Activities and the Establishment of a National Registry was adopted by the Austrian Parliament on 6 December 2011 (Marboe). It was necessitated by the approaching launch of two Austrian Cubesats, and addresses, like most other national space laws, mainly the issues of responsibility, authorization and supervision, liability and registration. It covers both activities in Austria as well as those carried out by Austrian nationals abroad. There is a mandatory insurance of 60 million Euros, and liability is limited to that amount, except in case of fault.

23.3 Exploration and Exploitation

The current legal regime governing exploration of outer space and celestial bodies is laid down specifically in the Outer Space Treaty and the 1979 Moon Agreement. As mentioned

above, the former has been ratified by 100 states and parts of it could be said to apply even to non-parties on the basis of having become customary international law. The latter has only 13 state-parties, none of which are established space powers (Australia, Austria, Belgium, Chile, Kazakhstan, Lebanon, Mexico, Morocco, the Netherlands, Pakistan, Peru, the Philippines, Uruguay). Even though the Moon Agreement has not been ratified as widely as the Outer Space Treaty, its relevance must not be underestimated. The Outer Space Treaty applies to outer space, including the Moon and other celestial bodies. The term 'celestial bodies' is however not defined. The Moon Agreement is a bit more precise; it applies to the Moon and other celestial bodies in the solar system other than the Earth, and reference to the Moon includes orbits around, or other trajectories to, or around, it. It does not apply to extraterrestrial materials that might reach the surface of the Earth by natural means.

Exploitation of lunar resources is considered to be the 'next step' in the conquest of space after exploration, and mainly the reason why the Moon Agreement has remained of limited influence to date. The Moon Agreement is the only one of the five UN space treaties that explicitly addresses exploitation, and discussions about the meaning of Article 11, declaring the Moon and its natural resources the "Common Heritage of Mankind," have sparked heated debate. The Moon Agreement prescribes that an international regime be set up to govern such exploitation, "as such exploitation is about to become feasible," and in relation herewith the question of the review of the Moon Agreement was foreseen 10 years after its entry into force, but this has never happened. The Moon Agreement entered into force in 1984, and no decision about review has since been taken—perhaps because exploitation is not yet quite around the corner. Despite the uncertainty about the exact implications of Article 11, it is clear that exploitation is not prohibited *per se* by the Agreement.

Both exploration and exploitation will be addressed in the following paragraphs. In terms of exploration, planetary protection will also be discussed, and claims to private property rights on the Moon and other celestial bodies will be included in the discussion regarding exploitation.

23.3.1 Exploration

General principles from the Outer Space Treaty governing Moon exploration and use include

- Freedom of scientific investigation.
- Province of all mankind.
- Non-appropriation.
- Compliance with international law including the UN Charter.

- Prohibition of nuclear weapons and weapons of mass destruction (not defined).
- International cooperation and mutual assistance.
- Non-interference with activities of other states.
- International (state) responsibility and liability, also for activities carried out by private entities (which require “authorization and continuing supervision”).

The Moon Agreement adds to this, including for instance

- Use for exclusively peaceful purposes.
- Prohibition of threats and hostile acts.
- Prohibition of military and weapons-related activities.
- Sharing of information on mission and its results.
- Report to the UN if discovery of organic life or phenomena endangering human life/health.
- Notification of placement or use of radioactive materials on celestial bodies.
- Any person on the Moon is considered an astronaut; refuge to be offered in case of distress.
- Non-interference and consultations for surface and underground activities/settlements.
- (Parts of) the surface or subsurface of the Moon, or natural resources “in place” may not become property of a state, IGO, NGO, national organization, non-governmental entity or natural person.
- Samples may be collected and removed for scientific purposes, appropriate quantities may be used to support missions.
- The Moon and its resources are the Common Heritage of Mankind and an international regime is to be established when exploitation of resources is about to become feasible.

As regards the important topic of the protection of the environment of celestial bodies, as indicated above, Article IX of the Outer Space Treaty provides a general obligation to protect all celestial bodies, including the Earth, from harmful contamination, which is not defined further. A similar provision is contained in Article 7 of the Moon Agreement, but it qualifies such contamination as taking place “through the introduction of extra-environmental matter or otherwise.” An IAA Cosmic Study was recently published on this subject and includes proposals such as a differentiation of space activities and areas of the Moon, a new interpretation of the term “due diligence,” the creation of “planetary parks” and a model for licensing procedures (IAA Cosmic Study on Protecting the Environment of Celestial Bodies, 2011).

Article 7 of the Moon Agreement also states the possibility of creating international scientific preserves for areas of the Moon having special scientific interests, thus providing an interesting means for protecting parts of the lunar environment for scientific research.

Currently attempts are being undertaken to ‘revive’ the Moon Agreement. Noteworthy is the 2008 Joint Statement

in the UNCOPUOS Legal Subcommittee by the states parties, attempting to convince other states to ratify the Treaty by highlighting its advantages, pointing out that in conjunction with the Outer Space Treaty, the Moon Agreement is helpful for rejecting “idle claims to property rights” that have surfaced in recent years. Also, the International Institute of Space Law (IISL) has issued two statements, in 2004 and 2008, about claims to private property rights in space (IISL Statements). The 2008 statement says

International Law establishes a number of unambiguous principles, according to which the exploration and use of outer space, including the Moon and other celestial bodies, is permitted for the benefit of mankind, but any purported attempt to claim ownership of any part of outer space, including the Moon and other celestial bodies, or authorization of such claims by national legislation, is forbidden as following from the explicit prohibition of appropriation, and consequently is prohibited and unlawful.

It is necessary to clarify and complement the legal regime currently regulating the exploration and use of the Moon and other celestial bodies. The broad principles that were adopted in the 1960s and 1970s remain valuable today and the delicate balance reached at that time should be maintained. However, additional regulation is necessary in order to ensure valuable, safe, economic, and fair exploration and exploitation that will benefit both current and future generations.

23.3.2 Exploration and Planetary Protection

‘Planetary protection’ is the term generally used to describe the guiding principles to be adhered to in the design of an interplanetary mission in such a manner as to prevent biological contamination of the target body, and, if appropriate, the Earth. These principles arise from the scientific need to preserve the target planetary conditions for future biological and organic constituent exploration. It also aims to protect the Earth and its biosphere from potential extraterrestrial sources of contamination in the event of a sample return mission.

Accordingly, a spacecraft must be sterilized before leaving Earth in order to minimize the risk of depositing terrestrial biological material at the target body. Any return vehicle must then be designed such that the sample is returned in a ‘contained’ manner with appropriate measures in place to dispose of any parts of the vehicle, which could have been contaminated before reentry into the Earth’s biosphere.

Clean room assembly and microbial reduction through heat, chemicals, or radiation are the basic techniques used to accomplish microbial control when this is necessary for a mission. These add a significant burden to mission

designers and integration teams. However, there is consensus that this is required in order to prohibit the possible microbial contamination of other planets. Article IX of the Outer Space Treaty states in relevant part that

States Parties to the Treaty shall pursue studies of outer space, including the Moon and other celestial bodies, and conduct exploration of them so as to avoid their harmful contamination and also adverse changes in the environment of the Earth resulting from the introduction of extra-terrestrial matter, and where necessary, shall adopt appropriate measures for this purpose.

The Committee on Space Research (COSPAR) has concerned itself with questions of biological contamination and space flight since its very inception, and maintains and promulgates planetary protection policy for the reference of spacefaring nations, both as an international standard on procedures to avoid organic-constituent and biological contamination in space exploration, and to provide accepted guidelines in this area to guide compliance with the wording of the 1967 Treaty and other relevant international agreements.

COSPAR classifications as applied to different planetary bodies can and will change due to new scientific knowledge. The discovery of extremophiles on Earth surviving temperatures that were previously thought to be lethal to all life demonstrate how difficult it can be to prevent biological contamination and set the appropriate levels of contamination and categorization.

23.3.2.1 COSPAR Policy

COSPAR recognizes that although the existence of life elsewhere in the solar system may be unlikely, the conduct of scientific investigations of possible extraterrestrial life forms, precursors, and remnants must not be jeopardized. In addition, the Earth must be protected from the potential hazard posed by extraterrestrial matter carried by a spacecraft returning from another celestial body. Therefore, certain space mission/target body combinations, controls on contamination shall be imposed, as introduced in [Chap. 17](#).

Assignment of categories for specific mission/body combinations is to be determined by the best multidisciplinary scientific advice. For new determinations not covered by this policy, such advice should be obtained through the auspices of the Member National Scientific Institutions of COSPAR. In case such advice is not available, COSPAR will consider providing such advice through an ad hoc multidisciplinary committee formed in consultation with its Member National Scientific Institutions and International Scientific Unions. The five categories for mission/target body type combinations and their respective suggested ranges of requirements are presented in [Table 17.4](#) and are described as follows

Category I includes any mission to a target body which is not of direct interest for understanding the process of chemical evolution or the origin of life. No protection of such bodies is warranted and no planetary protection requirements are imposed by the COSPAR policy. Examples of Category I missions include: Flyby, Orbiter, Lander: Venus; Moon; Undifferentiated, metamorphosed asteroids.

Category II missions comprise all types of missions to those target bodies where there is significant interest relative to the process of chemical evolution and the origin of life, but where there is only a remote chance that contamination carried by a spacecraft could jeopardize future exploration. The requirements are for simple documentation only. Preparation of a short planetary protection plan is required for these flight projects primarily to outline intended or potential impact targets, brief pre- and post-launch analyses detailing impact strategies, and a post-encounter and end-of-mission report which will provide the location of impact if such an event occurs. Solar system bodies considered to be classified as Category II include: Flyby, Orbiter, Lander: Comets; Carbonaceous Chondrite Asteroids; Jupiter; Saturn; Uranus; Neptune; Pluto/Charon; Kuiper-Belt Objects.

Category III missions comprise certain types of missions (mostly flyby and orbiter) to a target body of chemical evolution and/or origin of life interest or for which scientific opinion provides a significant chance of contamination which could jeopardize a future biological experiment. Requirements will consist of documentation (more involved than Category II) and some implementing procedures, including trajectory biasing, the use of clean rooms during spacecraft assembly and testing, and possibly bioburden reduction. Although no impact is intended for Category III missions, an inventory of bulk constituent organics is required if the probability of impact is significant. Solar system bodies considered to be classified as Category III include: Flyby, Orbiters: Mars; Europa.

Category IV missions comprise certain types of missions (mostly probe and lander) to a target body of chemical evolution and/or origin of life interest or for which scientific opinion provides a significant chance of contamination which could jeopardize future biological experiments. Requirements imposed include rather detailed documentation (more involved than Category III), including a bioassay to enumerate the bioburden, a probability of contamination analysis, an inventory of the bulk constituent organics and an increased number of implementing procedures. The implementing procedures required may include trajectory biasing, clean rooms, bioload reduction, possible partial sterilization of the direct contact hardware and a bioshield for that hardware. Generally, the requirements and compliance are similar to the Viking missions, with the exception of complete lander/probe sterilization. Category

IV specifications for selected solar system bodies are set forth in the Appendix to this document. Solar system bodies considered to be classified as Category IV include: Lander Missions: Mars; Europa.

Category V missions comprise all Earth-return missions. The concern for these missions is the protection of the terrestrial system, the Earth and the Moon. (The Moon must be protected from back contamination in order to retain freedom from planetary protection requirements on Earth-Moon travel.) For solar system bodies deemed by scientific opinion to have no indigenous life forms, a subcategory ‘unrestricted Earth return’ is defined. Missions in this subcategory have planetary protection requirements on the outbound phase only, corresponding to the category of that phase (typically Category I or II). For all other Category V missions, in a subcategory defined as ‘restricted Earth return’, the highest degree of concern is expressed by the absolute prohibition of destructive impact upon return, the need for containment throughout the return phase of all returned hardware which directly contacted the target body or unsterilized material from the body, and the need for containment of any unsterilized sample collected and returned to Earth. Post-mission, there is a need to conduct timely analyses of any unsterilized sample collected and returned to Earth, under strict containment, and using the most sensitive techniques. If any sign of the existence of an extraterrestrial replicating entity is found, the returned sample must remain contained unless treated by an effective sterilizing procedure. Category V concerns are reflected in requirements that encompass those of Category IV plus a continuing monitoring of project activities, studies and research (i.e. in sterilization procedures and containment techniques). ‘Restricted Earth return’ would include Mars and Europa, whereas ‘unrestricted Earth return’ would relate to the Moon.

23.3.2.2 Reporting

COSPAR recommends that its members provide information to COSPAR within a reasonable time (not to exceed six months after launch) about the procedures and computations used for planetary protection for each flight, and again within one year after the end of a solar system exploration mission about the areas of the target(s) which may have been subject to contamination. COSPAR maintains a repository of these reports, makes them available to the public, and annually delivers a record of these reports to the Secretary General of the United Nations.

The Reports should include, but not be limited to, the following information

- The estimated biological burden at launch, the methods used to obtain the estimate (e.g. assay techniques applied to spacecraft or a proxy), and the statistical uncertainty in the estimate.
- The probable composition (identification) of the biological burden for Category IV missions, and for Category V ‘restricted Earth return’ missions.
- Methods used to control the biological burden, decontaminate and/or sterilize the space flight hardware.
- The organic inventory of all impacting or landed spacecraft or spacecraft-components, for quantities exceeding 1 kg.
- Intended minimum distance from the surface of the target body for launched components, for those vehicles not intended to land on the body.
- Approximate orbital parameters, expected or realized, for any vehicle which is intended to be placed in orbit around a solar system body.
- For the end-of-mission, the disposition of the spacecraft and all of its major components, either in space or for landed components by position (or estimated position) on the surface of a celestial body.

23.3.3 Exploitation

The major challenge in creating a workable regime for the exploitation of outer space resources is to find the right balance between “benefit and interests of all countries” as proclaimed in Article I of the Outer Space Treaty. And, the equally vital need for return on investment and legal certainty for entrepreneurs—that need has also been explicitly recognised in the 1996 ‘Declaration on International Cooperation in the Exploration and Use of Outer Space for the Benefit and in the Interest of all States, Taking into Particular Account the Needs of Developing Countries’ (UN Res. 51/122, 1996, the so-called ‘Space Benefits’ Declaration). This Resolution, on the one hand, says that international cooperation in the exploration and use of outer space must take particular account of the needs of developing countries. While on the other hand, it recognizes that states are free to determine all aspects of their participation in such cooperation on an equitable and mutually acceptable basis, and that contractual terms should be fair and reasonable and should comply with the legitimate rights and interests of the parties concerned (e.g. intellectual property rights).

Parallels for the regime governing the exploration and exploitation of the Moon can be found in the Law of the Sea regime and in the Antarctica regime. The Law of the Sea regime also contains the term “Common Heritage of Mankind” with regard to resources of the deep seabed. Subsequent amendments have attempted to bring the system more in line with political and economic realities, and thus more readily acceptable by all states. As far as the Antarctic regime is concerned, the situation is somewhat different, as several states have claimed sovereign rights over the area,

which have subsequently been ‘frozen’ but which are still ‘around’ (this is not the case for the celestial bodies or parts thereof). In 1991 the ‘Consultative Parties’ (i.e. the most interested parties with regard to these claims) decided to refrain from mining Antarctica and to “commit themselves to the comprehensive protection of the Antarctic environment and dependent and associated ecosystems and hereby designate Antarctica as a natural reserve, devoted to peace and science”. The mineral resources of Antarctica have not been declared the “Common Heritage of Mankind” (PEX Report; Towards a Global Space exploration Program: a Stepping Stone Approach, COSPAR Panel on Exploration, 2010).

In the light of this, one may wonder whether it is necessary to ‘renegotiate’ or otherwise amend the Moon Agreement, in order to establish an ‘authority’, like in the Law of the Sea regime, for example, or to transform it into something more similar to the Antarctic Treaty System.

23.3.4 Exploitation and the Role of Private Enterprise

Having established that exploration and exploitation of celestial bodies and their resources are permitted, albeit on certain conditions, another question that arises is whether they may be carried out by entities other than the state, namely by private enterprise. Article VI of the Outer Space Treaty states, in relevant part, that states parties must assure that national activities are carried out in conformity with the provisions set forth in the present Treaty, and must carry out authorization and continuing supervision over the activities of non-governmental entities in outer space, including the Moon and other celestial bodies.

Article 11 of the Moon Agreement further specifies that neither the surface nor the subsurface of the Moon, nor any part thereof or any natural resources in place, shall become property of any state, international intergovernmental or non-governmental organization, national organization or non-governmental entity or of any natural person.

It is thus clear that as long as some criteria are met, private entities are allowed to carry out activities in outer space. These requirements are further refined in other instruments, such as the 1972 Liability Convention. A state (i.e. the ‘appropriate’ state—some controversy exists as to whether that means the state of incorporation, main place of business, or otherwise) must authorize and supervise the private entity, and remains internationally responsible for its activities. This is increasingly being considered to be best ensured by way of a national licensing scheme—even recognized by the UNCOPUOS, and as discussed above, more and more states have taken that advice to heart and have designed or are thinking about developing national legislation in that sense.

23.3.4.1 The Need for Private Enterprise

The observation that private entities are allowed to undertake space activities, including exploration and exploitation of outer space as long as they fulfill certain requirements is important because the role of private enterprise increases by the day. The space treaties were drafted at a time when private enterprise was still far away, or private companies acted as contractors for the government. Nowadays they are more often than not the primary players, although some states still consider that space activities should be carried out only by governmental entities.

This means that the space treaties, although they allow and regulate the basics for private enterprise, do not settle all details, such as the exact nature of the relations between government and private entities, or the relations between different private entities. These details need to be worked out, and this can be done in the spirit of, and without nullifying, the existing treaties. Private enterprise today has become essential for further exploration and exploitation, it has brought economies of scale and arguably has even assured the continuation of space activity, and must be encouraged and facilitated by further elaboration of the existing legal frameworks. This must also be seen against the backdrop of ever more and closer international cooperation, which entails further complexities and interrelationships—as embodied in the International Space Station (ISS).

The requirement for authorization and continuing supervision (Article VI of the Outer Space Treaty) provides a significant measure of protection to private entities. The obligation to prevent harmful interference (Article IX of the Outer Space Treaty) means that in the event of harmful interference, international consultations could be conducted, or liability could be imposed, and there even is a duty to take part in such consultation when interference occurs.

In addition, one level down from the international treaties, the positive effects of national licensing schemes must be reemphasized. They contribute to an atmosphere of increased clarity and certainty. A state that grants a license to a private entity is unlikely to interfere with a project that is being operated in a legal and lawful manner. The licensing state is also unlikely to license another private entity to directly interfere with a previously authorized mission, and can take actions to protect the interests of private entities operating under its license, although it could revoke a license, cancel a frequency assignment, etc. Intellectual property rights claims could be also brought, or claims against unfair competition (Masson-Zwaan, *Lunar exploration and exploitation* etc., 2008).

Regrettably, some recent projects that seek to make a profit at any cost are undermining the balance and consensus that has been reached over the past 40 years. They

argue that space law does not protect their interests, and even prevents them from making a profit. They are especially critical of Article II of the Outer Space Treaty and the non-appropriation principle. It is accepted that the non-appropriation principle of Article II also forbids any private claims. It has been argued that use of the term ‘national’ appropriation exempts private entities, and therefore permits so-called ‘private appropriation’, but to date no persuasive arguments have been offered to justify this interpretation. It is a firmly established principle of international law, recognized by all, that a state can never grant more authority to a non-governmental entity than it possesses itself. Thus, states cannot authorize and license their nationals to appropriate outer space and celestial bodies, which is prohibited to the state itself under international law. The IISL Statements cited above conclude that the prohibition of national appropriation by Article II includes appropriation by non-governmental entities. The 2008 Statement intentionally ends on a positive note, by recalling that private activities on the Moon and other celestial bodies are permitted as long as they are in accordance with international space law, and subject to the authorization and continuing supervision of the appropriate state party.

Moreover, private ownership of celestial property is not even necessary for commercial use of resources, even on Earth. A private entity can perfectly legally and profitably extract and use resources from property that it does not own: think of offshore oil platforms or fisheries zones.

The relations between government and private entities need to be further regulated. More national licensing schemes need to be elaborated and promoted in many more countries. If possible in a framework of international consultation and coordination, possibly under the umbrella of UNCOPUOS, and of course the space law community must continue and intensify its efforts to convince their own, as well as—or even more so—engineers, scientists and politicians, that law and order in space can promote and encourage progress, as long as it is developed wisely, in a concerted effort between all parties concerned, and continuously evaluated and adapted to changing needs of all stakeholders.

23.4 Private Commercial Human Space Flight

Technological novelties never stop challenging lawyers’ abilities to adapt themselves, and the law, to new and unforeseen situations. This is especially so with regard to activities taking place in the sky and beyond. In the beginning, a law for activities in the atmosphere (or air space) was formulated. Next, a law of outer space came into being. In the near future, a novelty will come about that

does not quite fit into either of these categories, or maybe considered to fall under both. This new activity is often referred to as (suborbital) space tourism. Is it aviation or space flight, or something new and hybrid that should perhaps be called aerospace flight? Are vehicles that will be used aircraft or spacecraft, or something new and hybrid, perhaps called aerospace craft?

23.4.1 Categories of Space Travel

There are basically three categories of private commercial human space travel, which are briefly outlined below, although the emphasis in the remainder of this chapter will be on the first of these (see Masson-Zwaan, *Article VI* etc., 2009).

23.4.1.1 Suborbital Flights

In suborbital space flight, orbital velocity is not achieved. After engine shutdown, 3–6 min of weightlessness is achieved, after which the vehicle falls back to Earth and reenters the atmosphere. Most current projects for private commercial human space flight will offer this kind of ‘space travel’. Vehicles usually attain a maximum altitude of around 100 km. Different technologies are under consideration. Some concepts involve a horizontal takeoff or launch (sometimes from an aircraft), while others take off vertically. For landing, they can vary from aircraft to parachute, the main technology challenge being thermal protection during reentry. Some see suborbital flights as equivalent to sounding rockets, not requiring licensing as either aircraft or spacecraft; however there is no (international) agreement yet as to the exact status of these flights, and the fact that they intend to carry commercial passengers and cargo certainly seems to differentiate them substantially from sounding rockets.

Perhaps the most renowned enterprise in this category is Virgin Galactic. The concept involves a launch of SpaceShipTwo in mid-air at 50,000 feet (>15 km) from the mothership, an aircraft called WhiteKnightTwo. Its home base will be Spaceport America in New Mexico, USA, but flights are also planned from other sites around the world. The 2½ h journey into space sells for US\$200,000 per seat, and hundreds of people have reportedly already signed up.

XCOR Aerospace is developing the two-seat Lynx suborbital spaceplane. The spaceplane will be capable of flying several times each day, with flights being retailed by Space Expedition Corporation (SXC). SXC may preempt Virgin Galactic in being the first to start commercial operations, possibly from Curaçao in the Caribbean, under a ‘wet-lease’ agreement with Space Expedition Curaçao. Tickets sell for US\$95,000.

23.4.1.2 Orbital Space Flight

In orbital space flight, orbital velocity must be achieved for the vehicle to keep flying along the curvature of the Earth and not fall back. Orbital space flight is technically highly complex and therefore expensive. Providing orbital space flight for private paying clients is much more demanding than suborbital flight, both in terms of technology and cost, but is nevertheless envisaged by several ventures, following in the footsteps of the seven private individuals who have flown on Russian spacecraft to the ISS. In April 2001, the first commercial space tourist Dennis Tito (born 1940) spent six days in the Russian section of the ISS, after extensive training at the Star City complex. Such flights are marketed by Space Adventures. The price for a flight to the ISS on board a Soyuz spacecraft is around US\$35 million. Seats have become scarce since the International Space Station doubled its crew size up to six people in May 2009, because there are now no spare seats available on the Soyuz flights.

23.4.1.3 Intercontinental Rocket Transport

Intercontinental rocket transport implies a transit through space in order to substantially shorten the travel time from one point on Earth to another. It is not a new idea, but the technical challenges are very demanding in terms of the velocity and the amount of propellant required, and the need for robust thermal protection for reentry. Cost is therefore prohibitive, at least in the short to medium term.

23.4.2 Realities

As several market studies have shown (e.g. Futron Space Tourism Market Studies, 2002, 2006), it is certain that space tourism will happen. The basic legal framework for private commercial space activity is in place, although the extent to which humankind might one day engage in commercial space tourism activities was not anticipated. More and more licensing systems are being put in place under national law, complementing the international legal framework, which will help to provide legal certainty and harmonized rules. Regulatory and legal certainty in this field is important for new industries but also for passengers and third parties.

Mass tourism is probably still several decades away. When ticket prices come down to US\$20–40,000, the numbers of passengers is likely to increase, as the prospect of experiencing weightlessness and observing the ‘Blue Planet’ from outer space is very attractive to many people. For the immediate future, it will be only for the rich few, at considerable personal risk, liability for which they will be requested to waive, while insurance may not yet be available.

23.4.3 Regulatory Needs

The multilateral space treaties elaborated within UNCOP-UOS were formulated in the Cold War era, when only a small number of countries had spacefaring capability. They could not fully anticipate the extent to which humankind would one day engage in commercial space tourism activities. The Outer Space Treaty foresaw that private entities would one day engage in space activities, yet one of the most essential topics for private operators, namely their exposure to second- or third-party liability, is not addressed. Second party or contractual liability refers to liability of the operator *vis-à-vis* passengers and cargo, while third-party or non-contractual (tort) liability refers to liability for damage to persons or property on the ground, who have no contractual relations with the activities of the operators.

Instead, the Treaty, as well as the Liability Convention, only addresses liability at the level of the states involved. There is no cap on liability of operators, and no opportunity for passengers or third parties to present direct claims for compensation. Thus, even though the Treaties maintain their relevance even after several decades, the existing international legal regime needs to be supplemented with additional and more specific rules. Once again, a balance must be found between commercial and technological opportunities on the one hand and principles of international space law on the other, and between the interests of the state and those of private enterprise and passengers and third parties. In essence, it is necessary to protect the legitimate interests of states and to ensure the safety of crew, passengers and third parties in a satisfactory way, without creating a regulatory overkill.

23.4.4 Does International Space Law or Air Law Apply to Space Tourism?

The UN space law treaties apply to relations between states in carrying out space activities. International air law conventions deal with international carriage by air.

Many of the currently planned space tourism projects plan to operate from one state, i.e. they will take off and land at one and the same spaceport. In these cases, the likelihood of international legal issues such as cross-border damage is limited, and in principle that state’s *national* law will apply. Most ventures are currently planned to take place in the USA. Here, no international element will be involved, and the state may determine whether it will regulate the activity as an aviation or space activity under its national law.

The USA has developed a substantive body of rules governing private human space flight. A ‘light touch’ legal approach has been taken, and licenses from the FAA’s Office of Commercial Space Transportation are mainly concerned with public safety, not so much with the safety of passengers (who are voluntarily engaging in a risky activity). However, there are other cases where the probability of an *international* element, and thus the applicability of international air or space law, is much less remote; for instance, when such a mission would launch from a site in Europe. Countries in Europe are much smaller and so the possibility of cross-border damage is greater. This could then lead to damage being caused by (the private entity of) one state to persons or property of another.

Even though some states in Europe have enacted national space legislation creating a licensing system, as discussed above, none of these contain specific rules on space tourism. The Dutch Act contains a provision stating that it can be declared “wholly or partly applicable to the organization of outer space activities by a natural or juridical person from within the Netherlands” (Sec. 2.2 b). The explanatory note specifies “This might include the commercial organization of space tourism activities.” Note that the Dutch Act does not apply to Curaçao (from where Space Expedition Corporation (SXC) will operate the XCOR Lynx). Curaçao became a separate state in 2010 and does not yet have a national space law. Sweden, which may become involved as a launch site in Kiruna, might treat a vehicle such as SpaceShipTwo as a sounding rocket, which under its current act is not regarded as a space activity.

Efforts at harmonization of space legislations in Europe seem to be barred by Article 189 of the Lisbon Treaty of 2009, and the risk is a patchwork of rules will result that may ultimately lead to flags of convenience and ‘forum shopping.’ Some representatives of the European Aviation Safety Agency (EASA), the European counterpart of the FAA, have expressed their view since 2008 that EASA is competent to regulate suborbital flights, but only for horizontal takeoff concepts such as SpaceShipTwo, and only for the part that takes place in the air. They have taken a less flexible approach than the FAA, basically considering these craft to be aircraft, calling them suborbital aeroplanes (SOA), and requiring full certification at the start of commercial operations (Marciacq e.a.). In the fall of 2011, the European Commission asked EASA to put all rule-making activities in the field of suborbital flights on hold, for budgetary reasons. This means that regulation of such flights will have to be regulated under national law—as long as no international element is involved in the course of the flight.

In the situation where an international element warrants the application of international law, the question is whether

air or space law would apply to damage. There are major differences in both regimes. In air law, passenger liability and the liability of the operator towards third parties on the ground is laid down in an elaborate system of rules that have been tested and clarified extensively by jurisprudence, while space law is based on a rudimentary state-based system of liability that has moreover never been put to the test in a court case.

Unfortunately, neither air law nor space law provides definite answers, either in international instruments or in national laws, about whether ‘space tourism’ could be covered by the respective legal regimes. At the international level, space law provides no definition or delimitation of space, nor a precise definition of spacecraft or space object, but it would appear to be relatively easy to conclude that space tourism could fall under the space law regime. In international air law, there is a definition of aircraft (“Any machine that can derive support in the atmosphere from the reaction of the air other than the reaction of the air against the earth’s surface”, see the ICAO Annexes). However, despite the technical definition discussed at the start of [Chap. 2](#) and referred to as the von Kármán ellipsoid, there is no legal definition of where airspace ends; to be covered by air law, the definition of aircraft may need to be reinterpreted to include suborbital space flights. Thus, it would appear that under current international or national air or space law there is no definite answer yet about the legal status of suborbital space tourism.

So, should air law apply for part of the journey and space law at some (as yet undefined) point during the activity? Is the case different for suborbital flights and for orbital flights? And, what about horizontal (aircraft) takeoff as opposed to vertical (rocket propelled) takeoff (or suborbital flights versus orbital flights)? The application of two totally different regimes to one suborbital flight may be the result; this would be both unsatisfactory and impractical. Of course, ideally, a comprehensive and uniform ‘hybrid’ legal regime of aerospace law should cover the complete launch and return journey of private individuals. Such a regime may well require a new Treaty, but the probability of states agreeing on a new Treaty is not high and might take decades.

Since there is a need to provide clarity to today’s entrepreneurs and to safeguard the interests of all players involved, an interim solution could be to apply space law to the entire suborbital flight, on the basis of the function of the vehicle or mission. Since the purpose of space tourism is to go to space, space law might be applied to the entire mission. But, appropriate clarifications and additions (perhaps based on the model of the US regulations) must be made to supplement the provisions of the space treaties, and

national legislation should address this activity as much as possible—as a case in point, more EU member states should enact national legislation in this field.

23.5 Conclusion

The general legal framework for space activities under public international law as contained in the UN treaties is in place, and is sufficiently general and flexible to enable and encourage states to carry out space activities in an orderly manner. It contains the basic provisions that allow states to address the legal implications of space activities. But it is also clear that the time has come for the international community to agree on the further development of these general principles, to deal with emerging issues such as exploration, planetary protection, exploitation, the interests of private entities or private human space flight; and certainly the problems related to space debris mitigation and remediation require attention. Such rules can of course be laid down in the form of treaties, but other forms such as national legislation, guidelines, codes of conduct and the like may be more efficient and realistic in the short to medium term. In creating this additional legal framework implementing the basic principles laid down in the UN space treaties, ideally in the form of guidelines and other ‘modern’ forms of international law, a continuous interaction between scientists, engineers and lawyers will remain of paramount importance.

23.6 Sources and Further Reading

On space law in general:

Texts of Treaties and Resolutions:

- <http://www.oosa.unvienna.org/oosa/en/SpaceLaw/index.html>

Space Law—A Treatise, F. Lyall and P. Larsen, Ashgate, 2009

On commercial issues:

- Contracting for Space, L.J. Smith and I. Baumann (eds.), Ashgate, 2011

On national space legislation:

List and texts of national space legislations:

- <http://www.unoosa.org/oosa/en/SpaceLaw/national/index.html>

Von der Dunk, F. (ed.), National space legislation in Europe, Brill, 2011

Gabrynowicz, J., One Half Century and Counting: The Evolution of U.S. National Space Law and Three Long-Term Emerging Issues, in 4 Harvard Law and Policy Review 405 (2010)

Rapp, L., When France puts its own stamp on the space law landscape, in 35 J. of Space Law 313 (2009)

Lazare, B., The French Space Operation Act: Technical Regulations, 2nd IAA Symposium on Private Human Access to Space, 2011 (on CD)

Marboe, I., Traunmüller, K., Austrian Federal Law on the Authorisation of Space Activities and the Establishment of a National Registry (Austrian Outer Space Act), presented at the Eilene Galloway Symposium on Topical Issues in Space Law, Washington DC, December 2011 (on file with the authors)

On exploration and planetary protection:

COSPAR documents:

- COSPAR RESOLUTION 26.5, *COSPAR Information Bulletin* 20, 25–26, 1964.
- COSPAR DECISION No. 16, *COSPAR Information Bulletin* 50, 15–16, 1969.
- COSPAR DECISION No. 9/76, *COSPAR Information Bulletin* 76, 14, 1976.
- COSPAR INTERNAL DECISION No. 7/84, Promulgated by COSPAR Letter 84/692-5.12-G. 18 July 1984, 1984.
- COSPAR DECISION No. 1/94, *COSPAR Information Bulletin* 131, 30, 1994.

DeVincenzi, D. L., P. D. Stabekis, and J. B. Barengoltz, A proposed new policy for planetary protection, *Adv. Space Res.* 3, #8, 13, 1983.

DeVincenzi, D. L., P. D. Stabekis and J. Barengoltz, Refinement of planetary protection policy for Mars missions, *Adv. Space Res.* 18, #1/2, 314, 1994.

Rummel, J. D., et al. Report of the COSPAR/IAU Workshop on Planetary Protection, COSPAR, Paris, France, 2002.

Space Studies Board, National Research Council (US), *Evaluating the Biological Potential in Samples Returned from Planetary Satellites and Small Solar System Bodies*, Task Group on Sample Return From Small Solar System Bodies, National Academy of Sciences, Washington, D.C., 1998.

Space Studies Board, National Research Council (US), *Preventing the Forward Contamination of Europa*. Task Group on the Forward Contamination of Europa, National Academy of Sciences, Washington, D.C., 2000.

IAA Cosmic Study on Protecting the Environment of Celestial Bodies (PECB), 2011.

On exploitation and private property rights:

- IISL Statements:

- Statement by the Board of Directors of the IISL on Claims to Property Rights Regarding the Moon and other Celestial Bodies (2004), and Further Statement by the Board of Directors of the IISL on Claims to Lunar Property Rights (2009), <http://iislweb.org/publications.html>

- PEX Report: Toward a Global Space Exploration Program: A Stepping Stone Approach, COSPAR Panel on Exploration (PEX), 2010
 - Masson-Zwaan, T., *Lunar exploration and exploitation as a special case of planetary exploration: legal issues*, in: Contemporary problems of international space law, A. Kapustin and G. Zhukov (eds.), 159 (Moscow, 2008)
- On private commercial human space flight:
- Masson-Zwaan, T. and Moro-Aguilar, R., *Practical Solutions for the Regulation of Private Human Sub-Orbital Flight: a Critical Analysis*, 2nd IAA Symposium on Private Human Access to Space, 2011 (on CD)
- Masson-Zwaan, T. & Freeland, S. *Between Heaven and Earth: the legal challenges of human space travel*, 66 Acta Astronautica 1597–1607 (2010)
- Masson-Zwaan, T., Regulation of Sub-orbital Space Tourism in Europe: A Role for EU/EASA? in 35 Air and Space Law nr. 3, 263–272 (2010)
- Masson-Zwaan, T., *Article VI of the Outer Space Treaty and Private Human Access to Space*, Proceedings of the International Institute of Space Law 2008, 536–546 (2009)
- Von der Dunk, F., Passing the Buck to Rogers: International Liability Issues in Private Space Flight, 86 Nebraska Law Review 400 (2007)
- Hobe, S., *Legal Aspects of Space Tourism*, 86 Nebraska Law Review 439 (2007)
- Freeland, S. Up, up and ... Back: The Emergence of Space Tourism and its Impact on the International Law of Outer Space, 6 Chicago Journal of International Law 1 (2005).
- Marciacq, J.B., et al. *Accommodating sub-orbital flights into the EASA regulatory system* www.congrex.nl/08a11/presentations/day1_S09/S09_05_Marciacq.pdf
- Futron market studies:
- Space Tourism Market Study (2002), <http://www.spaceportassociates.com/pdf/tourism.pdf>
 - Suborbital Space Tourism Demand Revisited (2006), http://www.futron.com/upload/wysiwyg/Resources/Whitepapers/Suborbital_Space_Tourism_Revisited_0806.pdf

Les Johnson and Jack Mulqueen

Before there is a funded space mission, there must be a clear and present need for the mission. Space science and exploration are expensive, and without a well-defined and justifiable need, no one is going to commit significant funding for any space endeavor. However, as discussed in [Chap. 1](#), applications of space technology are many and broad, hence there are many ways to determine and establish a mission need.

Robotic science missions are justified by their science return. To be selected for flight, questions like these must be addressed: What is the principal science question that needs answering, and will the proposed mission be the most cost-effective way to answer it? Why does answering the question require an expensive space flight, instead of some ground-based alternative? If the question can only be answered by flying in space, then why is this approach better than other potential approaches? How much will it cost? And is the technology required to answer the question in-hand and ready to use? If not, then how much will it cost and how long will it take to mature the technology to a usable level?

There are also many ways to justify human exploration missions, including science return, technology advancement, as well as intangible reasons, such as ‘national pride’. Nonetheless, many of the questions that need answering are similar to those for robotic science missions: Where are the people going, why, and will the proposed mission be the most cost-effective way to get there? What is the safest method to achieve the goal? How much will it cost? And is the technology required to get there and keep the crew alive in-hand and ready to use? If not, then how much will it cost and how long will it take to mature the technology to a usable level?

Another reason for some groups sending spacecraft into space is for profit. Telecommunications, geospatial

imaging, and tourism are examples of proven, market-driven space missions and applications. For this specific set of users, the outstanding questions include: What is the product or service? Who will buy it? How can it be profitable? What is the most cost-effective solution to fielding the product or service? And, of course, is the technology in-hand or is further development required?

In order to answer these questions, the responsibility falls to a specially skilled set of engineers and scientists who understand how to assess the readiness of new technologies. This is a process of defining preliminary mission requirements, and the methodologies for assessing multiple candidate mission implementation scenarios against each other to achieve a point design for cost assessment, management review, and sometimes approval to proceed with development. This chapter will describe and discuss these advanced concept assessments.

24.1 An Advanced Concepts Team

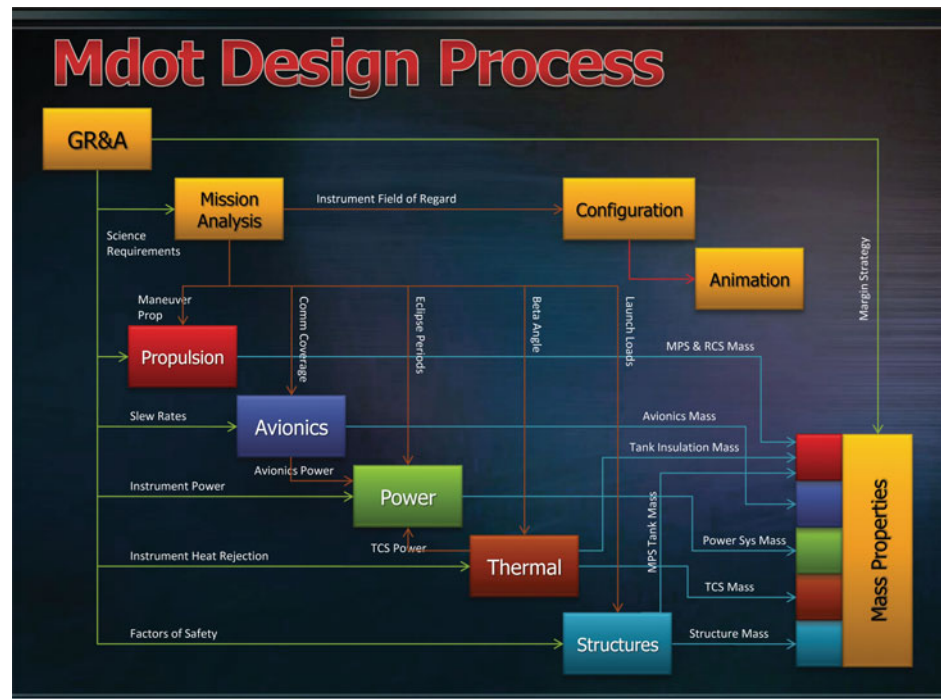
The specific skill mix and organization structure of an advanced concepts analysis team will vary with separate organizations. Some organizations will have a dedicated team of discipline engineers skilled in making high level, rapid turnaround concept studies funded and available when new analyses are required. Other organizations maintain only a core set of advanced concept managers, with discipline expertise obtained by ‘buying it’ from elsewhere within the organization, as and when specific studies need to be performed. Both approaches have been proven successful; the key is the attitude and training of the individual team members.

A successful space system advanced concepts analysis team will have experts in the following fields engaged in studies, as their skills are needed

- *Study Manager*—The primary interface with the customer or the innovator. This is the person that understands the

L. Johnson (✉) · J. Mulqueen
George C. Marshall Space Flight Center, National Aeronautics
and Space Administration (NASA), Huntsville, AL, USA
e-mail: C.Les.Johnson@nasa.gov

Fig. 24.1 The advanced concepts process used by the NASA Marshall Space Flight Center



requirements and can turn these requirements into a study plan that includes a schedule with its interim and final study products clearly defined. The Study Manager also develops the budget and staffs the team with the appropriate discipline experts for the assigned task. This person is also responsible for documenting the results of the analysis and providing it to the customer. It is also desirable for this person to report the results at a relevant technical conference or in a journal article, as appropriate.

- **Lead Systems Engineer**—This is ideally an experienced engineer who has seen at least one advanced concept from the idea phase through hardware development or space flight. It is their responsibility to make sure that the study products are wholly integrated, that they are internally consistent with all the study's ground rules and assumptions, and that there are no unforeseen system-level impacts resulting from any single team member's technical analysis.
- **Discipline Engineers**—These are the engineers who actually perform the technical analysis that is the basis of the advanced concepts study. Disciplines often represented in studies include power, avionics, thermal, configuration and layout, structures, mass properties, trajectory analysis, operations, propulsion and attitude control, human factors (optional and unique to human space flight concepts), cost, and risk.

24.2 The Advanced Concepts Process

The Mdot process used by the NASA Marshall Space Flight Center's Advanced Concepts Office for performing an advanced concept analysis of a potential future space mission or system is shown in Fig. 24.1. 'Mdot' is derived from the rocket equation—the mass flow rate.

The process usually begins with a thorough definition of the study's ground rules and assumptions (GR&A). It is in this phase that the customer describes the concept to be assessed or used as the basis for a mission concept definition. A cautionary note is in order: Many customers have preconceived ideas regarding how their technology might or might not be fielded and how it should or should not be implemented. Unless there is great care in the definition of the GR&A, these notions—which are really nothing more than notional engineering design solutions—might get listed as either a ground rule or an assumption. This must be avoided. It is up to the advanced concept study team to address these issues and to back them up with detailed engineering analysis. Poor GR&A can ruin a concept study.

The next step is usually a first cut mission analysis, or trajectory identification, based on gross payload mass and customer-provided destination requirements. This is done to bound the problem and to make sure that it is even possible for the spacecraft under consideration to meet a customer's

requirements. From this will flow general propulsion and attitude control system requirements in the form of an overall required velocity increment (ΔV). For Earth orbital missions, it is here that overall end-of-life deorbit propulsion requirements, if any, will be identified.

The propulsion system is sized to perform all of the orbital maneuvers required during the mission. These maneuvers may include initial orbit insertion, orbit altitude changes, orbit plane changes, orbit altitude maintenance. In many cases the propulsion system will consist of a main propulsion system (MPS) to perform orbital maneuvers and an auxiliary reaction control system (RCS) to provide spacecraft attitude control. The design of the propulsion system depends on many parameters in addition to the mission ΔV budget. The spacecraft mass usually drives the propulsion system thrust requirements, and either the mission duration or the number of maneuvers usually drives the propellant selection.

The avionics subsystem analysis includes the sizing of the data storage system, communications system, and control system. The avionics subsystem design is usually driven by the data quantity that is to be transmitted, the spacecraft distance from Earth, the required communications network, and the pointing and slew rate requirements for the spacecraft concept. One of the driving factors in sizing a spacecraft's power system is communications. If a mission is to operate far from Earth, or requires very large data rates, then this could be a significant design driver for the spacecraft's power system. It may also dictate pointing requirements, especially for missions deep in the outer solar system that require large communications antennas on the spacecraft.

The power system design is based on the electrical power requirements of the spacecraft subsystems, including the avionics, propulsion system controllers, and the heaters that are required in order to maintain proper temperatures for spacecraft systems. For human missions, power must also be supplied to life support systems. Electrical power is usually produced by solar arrays attached to the spacecraft. The amount of power generated is dependent on the size of the solar arrays and the angle between the Sun and the solar array, which usually depends on the orbital plane of the spacecraft. The power system design is often complicated by the fact that a spacecraft in low Earth orbit will spend considerable time in the shadow of the Earth. The power system must therefore be sized to produce more than enough power during the sunlit portions of the orbit and send the excess power to an electrical storage system capable of delivering power when the spacecraft is deprived solar power.

The thermal control subsystem design is based on a comprehensive analysis of the thermal balance of the environment; i.e. the heat generated by spacecraft systems

and the temperature requirements of the spacecraft. The spacecraft must be shielded from the radiated heat of the Sun and insulated from the extreme cold in shadows. The heat generated by the spacecraft electrical system must also be dissipated in order to maintain proper operation temperatures. The thermal control system usually consists of thermal insulation covering the external surfaces of the spacecraft, electrical heaters to maintain equipment temperatures, and thermal radiators to prevent excess heat build-up.

The spacecraft structural design is based on the configuration requirements of the spacecraft and analysis of the structural loads placed on the spacecraft. The configuration is dependent on many factors, such as placement of scientific sensors, placement of solar arrays, propulsion system size, and in the case of human missions, and the crew system layout. In many cases, the most significant configuration driver is the packaging of the spacecraft in the launch vehicle payload shroud. The structural loads are usually greatest during the launch and ascent to orbit. The spacecraft structures must have sufficient strength and stiffness to withstand the acceleration loads and vibrations during launch.

24.3 "I Have an Idea!"

The first essential step for assessing a new concept is to answer the question: What is the need? Many technologists are so enamored with their innovation that they fail to understand that no one will support it if it doesn't meet someone's needs. It is best to discuss or describe the innovation by its functionality and mission-level impact taking into account as many anticipated system-level impacts as possible—as identified in a thorough advanced concepts analysis. For example, a new technology for producing abundant power in deep space seems like the kind of innovation that would be of interest to anyone considering missions into deep space. However, for many robotic science missions it is not necessarily advantageous to have more power since there may not be any science instruments with such a requirement. If an entirely new paradigm, infrastructure, and instrument technology base is required to use the new power source then it may not be cost effective to implement, even if a potential customer were to fully appreciate how it might benefit their research.

In the case of a new power system that significantly increases the power availability in deep space, an advanced concepts analysis would also have to be performed to fully understand the system level impact of the new technology on the rest of the spacecraft. Some questions to be asked include

- How will the spacecraft get rid of the extra heat load?

- Typically, more radiators will be required, increasing both the weight and cost of the spacecraft, not to mention its increased complexity.
- Will any spacecraft and payload science instruments consider the new power source as a new source of background noise?
 - Science instruments can be sensitive to background electromagnetic (EM) radiation, in which case they may be adversely effected by the additional EM radiation from the new, high power system.
- Are there safety issues with launching the new power source?
 - Any sufficiently compact, high-density power source that is miniaturized to fit within a spacecraft is only a small step away from being an explosive with clearly associated mission risks.

24.4 “How Do I Get It Selected for Flight?”

There are many good ideas out there for space missions, whether they are in space science, exploration, or advanced technology development. Unfortunately, there is always limited funding. The shortage of money therefore drives the bureaucratic system of most governments into having a standard processes by which missions and technologies are selected for flight. Learning these processes is vital to the advanced concept advocate.

Most people think that winning a flight happens as the result of writing a good proposal once a government solicitation is released. While this is strictly true, it does not tell the whole story. In fact, while a good proposal is necessary to win, it is, by itself, woefully insufficient. Most proposals are actually ‘sold’ before the formal proposal is ever written. To be successful, the advocates should consider a variation to this plan of action before the date of a solicitation is even announced

1. Make your idea widely known by presenting it at technical conferences and in journals.
2. Attend the discipline-specific meetings that are held by the potential customers and advocates, presenting your idea, even if the conference’s topic is not something directly related to your primary interests. An example might be a small spacecraft manufacturer who has conducted an internal study of using a new attitude control algorithm for their spacecraft bus attending a gathering of solar physicists because the manufacturer knows that such a pointing stability will be of significance in future solar physics missions.
3. Find out who the deciding official will be for an expected procurement, and go visit with them, discussing your great new idea, before the procurement is

ever released. Your goal is to influence the procurement so that your idea is absolutely within its scope.

4. Find other potential users, even those who may not have any money to fund it at this time, and get letters of interest or support for use in upcoming competitive solicitations.
5. Repeat steps 1 to 4, as necessary, until the selection and funding of your idea.
6. In parallel to the above, partner with industry, academia, or even a government agency to broaden the political and technical support for your idea. Having internal champions within the sponsoring organization significantly increases your odds of being selected.
7. With your step 6 partners, complete a high-level mission concept study that will allow you to have graphic images or even artist depictions of your idea. A picture is worth a thousand words, but an engineering drawing is worth at least getting the technological readiness level to TRL-6 with a shot at TRL-7.
8. Don’t oversell. Be honest in your trade studies when it comes to the pros and cons of your ideas versus the competition. Just make sure you highlight the pros and have a ‘good answer’ to the cons—good, in the sense that you have a plan to attack whatever the problem may be.
9. When the solicitation is released, don’t go after it alone. Yes, you and your organization may be the best people in the world to do the work, but partnering with others provides enhanced advocacy and a sense of the idea’s importance.
10. Get ready to lose; but in the loss, find out from the reviewers what they deemed needed improvement so that you will become better prepared for the next opportunity.

24.5 Crossing the TRL Valley of Death

As discussed in [Chap. 21](#), the problem of insufficient technical readiness can prevent missions from using new technologies, thus reducing potential returns, and the subsequent entrapment of new technologies without sufficient flight validation to reduce their inherent risk—potentially ‘forever’ preventing the new technology or approach from being selected for a flight mission. Many technologies find themselves at this critical juncture, known as the TRL ‘valley of death’, because they are too advanced for further ground-based research and development, yet have been insufficiently proven to be accepted for a flagship science or exploration mission because they have never before been proven in space.

The ‘valley of death’ exists because of the inherent high cost of flying missions in space. The cost of maturing most space technologies from one TRL to the next is relatively

inexpensive when compared to the cost of going from TRL-6 to 7. In fact, for many technologies the cost of going this last step is far more than all the money spent to take them from TRL-1 to TRL-6 combined.

24.6 Advanced Concepts Analysis in Technology Selection

A good advanced concepts analysis should result in a spacecraft or vehicle concept that will eventually be proven to have been within 30 % of its eventual mass and cost. While not a detailed design, concept analysis will nonetheless provide a configuration, mission scenario, spacecraft or vehicle configuration, mass and power budgets, materials list and integrated mass table (with margin), and a candidate launch vehicle capable of lofting the payload to its desired destination.

Within aerospace generally, advanced concepts analysis is used in a wide range of areas, including the following examples. *Future Space Missions*: Mission design includes defining outcomes, designing for the mission environment, planning for mission ground support, and considering follow-on missions. As with spacecraft design, mission design must consider end-to-end planning, from the initial funding to the system's retirement: system costs, operational needs, hardware and software interactions, mid-mission problem-solving, and hardware disposal. Advanced concepts analysis digs down to the component level of design, but also takes the '50,000-foot' view to ensure that a human or robotic mission operates in the way it was intended. *Space Transportation Concepts*: Starting from the ground up, space transportation systems must be considered from liftoff, to in-space operations, to atmospheric entry. When advanced concepts conducts planning for space transportation systems, all aspects of the work must be considered, from propellant use to propulsion system mass and performance to payload interfaces. This sort of preliminary planning ensures that hardware traveling into and through space is optimal for its intended mission, and that it can function properly when it arrives at its destination. *Launch Vehicle Concept Design*: Launch vehicles are defined to ensure that payloads of a specific weight reach the proper altitude above the Earth. A thorough definition process will include reviews of current, in-work, and theoretical designs for space missions to ensure that a launch vehicle design is optimized to meet a particular class, or classes of mission needs. Using the industry standard and organization unique models (when they exist), the analysis should evaluate the safety and success of the vehicles that take space missions from the ground into space. *Integrated Space Systems Analysis and Design*: Whether it be a life support system, a vehicle, or a mission requiring multiple pieces of hardware

and software, analysis of the interactions between multiple systems and subsystems will help mission planners make informed decisions about future designs. The analysis should anticipate problems before they appear in the hardware, thereby allowing for major incompatibilities to be remedied in the relatively inexpensive concept definition and design phase of a project.

24.7 Conclusion

A successful advanced concept analysis team looks like a miniature engineering organization in terms of skill mix, and like an integrated product team in terms of staffing. It must be small, experienced in working at a fairly high level (in other words, not so detail oriented as to preclude the ability to produce rapid turnaround engineering analyses with accuracies of about a factor of two), and able to iterate many different concept design options rapidly.

Team analysis must begin by gaining a thorough understanding of a mission's needs, and establish a close working relationship with the customer to make sure the final concept is aligned with both their stated and unstated requirements.

Finally, the team must have enough project experience to understand the difference between paper-study feasibility and engineering capability. TRL is one of the tools that can be used to make this assessment. However, there is no substitute for experience, and having a team populated with engineers who have worked on a successful hardware project in the past is a definite advantage, and tends to produce a more realistic advanced concept design.

24.8 Case Study

There are numerous examples of successful advanced concepts studies available in the literature. One led by the authors of this chapter is provided herein.

24.8.1 Integrated In-Space Transportation Plan

Advanced In-Space Propulsion (ISP) technologies will enable much more effective exploration of our solar system, and will permit mission designers to plan missions to 'fly anytime, anywhere and complete a host of science objectives at the destinations' with greater reliability and safety. When compared with state-of-the-art chemical propulsion, increased capabilities include shorter trip times to outer planets, higher payload mass, and enabling of missions that are either very difficult or impossible with chemical propulsion. Examples of these missions are orbits around the

outer planets, interstellar probes, and sample return missions from Mars or other planets. With a wide range of possible missions and many candidate propulsion technologies with very diverse characteristics, the question of which technologies are 'best' for future missions, is a difficult one. Resource limitations do not permit the development of all candidate propulsion technologies. Therefore, it is required to develop a set of propulsion technologies that will adequately satisfy a broad spectrum of mission requirements.

In the early 2000s, NASA tasked the NASA Marshall Space Flight Center to lead a national effort to identify promising ISP technologies, assess their ultimate capability to perform various future science and exploration missions, and recommend which should be funded for further development.

The effort was broken down into five parts: (1) address missions, mission priorities, and mission requirements as defined by the various NASA mission directorates; (2) provide a forum for technologists to advocate any ISP technology for any mission(s) for which they deemed their propulsion technology to be appropriate; (3) perform system analyses of the prioritized mission set to the degree necessary to support evaluation and prioritization of each technology advocated by the technologists; (4) perform cost analyses on each of the technologies that were determined by systems analyses to be viable candidates for the mission set; and (5) integrate all customer, technologist, systems, cost, and program inputs into a final prioritized set of technologies.

The primary products were a prioritized set of advanced ISP technologies that meet customer-provided requirements for the customer prioritized mission set and a set of recommendations of the relative technology payoffs in order to guide future NASA investment decisions. This effort involved many people at most NASA centers. The effort was divided among several teams

- The missions requirements team (MRT) defined the missions of interest and established the requirements for each.
- The systems team (ST) performed systems analyses to derive the important mission parameters for each propulsion technology for each mission. It also scored each technology for each mission against the figures of merit for performance, technical characteristics, and reliability/safety. The team consisted of 25 people from six NASA centers and three private companies.
- The Technology team (TT) proposed candidate propulsion technologies to be applied to each of the missions and provided the important performance and technical characteristics for each of the proposed technologies. It also performed scoring for figures of merit related to

Table 24.1 Future NASA missions as high priority candidates for new in-space propulsion technologies

Mission category	Missions of interest
Earth vicinity, low to moderate delta velocity (ΔV)	Geospace electrodynamic connection (GEC)
	Low earth orbit synthetic aperture radar (LEO SAR)
	Natural haz. and soil moisture measurement SAR
	Earth radiative energy meas. facility (Leonardo)
	Magnetospheric constellation (MC)
Inner solar system, simple profile, moderate ΔV	Ionospheric mappers
	Space interferometry mission (SIM)
Inner solar system, sample return	StarLight ST-3
	Comet nucleus sample return (CNSR)
Inner solar system, complex profile, moderate to high ΔV	Mars sample return (MSR)
	Earth atmospheric solar occultation imager (EASI)
	Pole-sitter (PS)
	Sub L1 point mission
	Solar sentinels
	Solar polar imager (SPI)
	Next generation space telescope (NGST)
Outer solar system, simple profile, high ΔV	Terrestrial planet finder (TPF)
	Outer zodiacal transfer
	Outer zodiacal transfer
Outer solar system, complex profile	Titan explorer (TE) (Titan organics orbiter/lander)
	Neptune orbiter (NO)
	Europa lander (EL)
	Solar probe
Beyond outer solar system	Interstellar probe (ISP)
Human visit to lunar, cislunar, and earth vicinity	Moon and earth-moon libration points
	Sun-earth libration points
Human visit to asteroids/Mars vicinity	Near-earth asteroids
	Mars piloted (MP) and cargo

schedule. The team consisted of 22 people from five NASA centers and two private companies.

- The cost team (CT) performed cost analyses and performed scoring on figures of merit related to cost. It consisted of four people from two NASA centers and two private companies.

- An advisory group (AG) performed oversight for the entire process. The group reviewed the mission selection, reviewed the figure of merit dictionary, set weights for figures of merit within each figure of merit category, set weights among the figure of merit categories, and performed the final prioritization from the data derived and presented. The group consisted of nine people from NASA headquarters and three NASA centers.

The MRT identified 28 missions of interest to NASA. These were allocated to one of nine different categories, according to mission destination and propulsion function at the destination (see Table 24.1). As available time and resources did not permit detailed analyses of all 28 missions, nine were selected on the basis of

- Missions rated as highest priority by the MRT.
- Maturity and completeness of mission requirements.
- Importance of availability of advanced propulsion technologies to the efficacy of the mission.
- Attainment of a representative set over a diverse range of mission requirements.

To ensure that the highest priority missions were analyzed first, the MRT prioritized missions within each mission category; italics in Table 24.1 denote the nine missions analyzed. For each mission analyzed, the top-level mission requirements were documented and maintained in a requirements document.

The study identified aerocapture, 5–10 kW solar electric ion propulsion, and nuclear electric propulsion as high priority technologies. Solar sails, 100 kW solar electric Hall-effect thrusters, and advanced chemical propulsion were identified as medium priority technologies. Plasma sails, momentum exchange tethers, and low density solar sails were identified as high risk/high payoff technologies primarily due to their relatively low technical maturity.

The results were used to prioritize the investments of the 200 million USD NASA In-Space Propulsion Technology Project from 2002 to 2005 that resulted in the successful maturation of aerocapture, 5–10 kW solar electric ion propulsion, and solar sail technologies to TRL 5/6 through extensive ground technology demonstrations.

Further Reading

1. Eberle, B., Farris, B., Johnson, L., Jones, J., Kos, L., and Woodcock, G., *Selection and Prioritization of Advanced Propulsion Technologies for Future Space Missions*, 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 7–10 July, 2002, Indianapolis, IN, USA.
2. Farris, B., Eberle, B., Woodcock, G., and Negast, B., *Integrated In-Space Transportation Plan*, NASA/CR-2002-212050, NASA George C. Marshall Space Flight Center, 2002.

Massimiliano Vasile, Stephen Kemble, Andrea Santovincenzo
and Mark Taylor

This chapter presents different approaches to the design of space missions and in particular the overall integration of systems and mission design. The chapter will start with the relationship between mission analysis and system design and the role of mission analysts in the context of the overall design process. It then continues with a section on the application of concurrent engineering principles to the design of space missions. The section starts by giving some definitions and a historical perspective on the use of concurrent engineering in the space sector and then illustrates the design process and the major actors and components. It provides the view of the European Space Agency on the preliminary design of space missions. The subsequent section provides a different perspective, namely the design of low-cost missions as seen from a private company.

The last section of this chapter presents a possible future development that imports into the design of space systems the principles of optimization, robust design, and design for reliability.

M. Vasile (✉)

Advanced Space Concepts Laboratory, Strathclyde Space
Institute, University of Strathclyde, Glasgow, Scotland
e-mail: massimiliano.vasile@strath.ac.uk

S. Kemble

Airbus Defense & Space, EADS Astrium, Stevenage, England

A. Santovincenzo

European Space Research and Technology Centre, European
Space Agency (ESA-ESTEC), Noordwijk, Netherlands

M. Taylor

Surrey Satellite Technology Ltd. (SSTL), Surrey, England

M. Macdonald and V. Badescu (eds.), *The International Handbook
of Space Technology*, Springer Praxis Books, DOI: 10.1007/978-3-642-41101-4_25,
© Springer-Verlag Berlin Heidelberg 2014

25.1 Mission Analysis and System Design

The term mission analysis generally refers to the area of analysis devoted to the orbital design and maneuvering aspects of a space mission. As such, a strong interaction exists between mission analysis and the analysis of core mission requirements (particularly the orbital coverage and environmental aspects). A further interaction exists with several spacecraft design aspects, including propulsion, communications, thermal, and power. These aspects are influenced by both the operational orbit and intermediate transfer orbit designs.

The activities within mission analysis can be broken down into a number of actions, each related to aspects of the mission requirements or spacecraft design. This section will describe, in compact form, the fundamental tasks involved in mission analysis and some of the techniques that can be used to implement them. [Section 25.2](#) discusses some of the issues involved in the design of certain types of operational orbits that are frequently encountered. References are provided from which more detailed information can be obtained.

25.1.1 Mission Analysis Tasks

Any organization procuring a satellite will have a set of requirements for the mission. These relate to both the design of the satellite and its payload, and also the orbital design aspects. Furthermore, there is an interaction between satellite design and orbital design. Orbit aspects determine the environment in which the satellite will operate, and as such influence the design of the communications, power, thermal and propulsion systems.

Mission requirements generally allow a degree of flexibility in the operational and transfer orbit design. Therefore, the task of a mission analyst is to design these elements in a way that enables an efficient satellite design. The process is

strongly interactive, involving satellite systems and sub-systems engineers.

The different elements of the mission analysis task can be summarized as follows.

- Design of operational orbits whose characteristics fulfill customer requirements, e.g. providing adequate coverage of sections of the Earth for an Earth observing mission, or providing global scans of a distant moon for an inter-planetary mission.
- Design of transfer orbits between launch injection orbit and the operational orbit, and the derivation of the associated maneuvers.
- Deriving satellite system related information arising from both the operational and transfer orbit design. These are dependent on a number of factors such as
 - Operational and transfer orbit environments, e.g. radiation and atmospheric effects.
 - Operational and transfer orbit geometry and kinematics, e.g. ground station contact, eclipse and general orbit illumination effects.
 - Operational and transfer orbit maneuver requirements, e.g. derivation of effects of satellite propulsion on the maneuver implementation and its efficiency.

In order to complete these tasks a number of key methods and techniques are frequently used. These are described in [Chap. 4](#).

The activities involved in mission analysis are closely connected with a number of spacecraft system design tasks. In some cases an iterative procedure must be adopted to simultaneously evolve the mission and subsystem designs in order to provide an efficient, combined solution. Efficiency is often related to minimizing total mass. Low mass satellites and systems are generally sought, whilst meeting the requirements specified by the customer.

25.1.1.1 Propulsion

Orbital maneuver requirements drive the propulsion design. Conversely, limitations imposed on propulsion system performance restrict the scope of maneuvers that may be performed, with implications for orbit choice, transfer duration and/or fuel requirement.

As discussed in [Chap. 11](#), the key parameters are thrust and specific impulse. Thrust and mass determine the extent of the orbital arc over which a maneuver must be performed in order to achieve the desired change, with implications for efficiency and transfer duration. Specific impulse determines the fuel required for maneuvers. However, high specific impulse often has a penalty attached, such as only being available with low thrust and making demands on spacecraft power. Therefore, propulsion and mission design can often be an iterative process to find an efficient solution

that meets all requirements. This can in some cases be treated as a constrained optimization problem, with the objective to maximize the payload or minimize launch mass.

25.1.1.2 Communications

Spacecraft distance, elevation above the ground station's horizon, and duration of the visibility pass all influence communications system design. Antenna type, amplifier type, power requirements, and even coding choice (see [Chap. 14](#)) can be linked to the mission design. Conversely, imposing restrictions on the type of communications units that may be used can significantly influence the choice of, in particular, the operational orbit. As in the case of propulsion, iteration is possible between orbit and system choices.

25.1.1.3 Power and Thermal Systems

The relationship between the orbit geometry and the Sun, together with the satellite pointing requirements to fulfill the overall mission needs (e.g. a nadir pointing system for Earth observation) influence the design of both the power and thermal systems; see [Chaps. 10](#) and [13](#), respectively.

Solar array size and also the degrees of freedom required of the array steering mechanism are influenced by orbit choice. Eclipse can be encountered, and this in turn influences the battery requirements. Solar illumination and eclipse also effect the type and sizing of the thermal system design.

Both operational and transfer orbits must be considered. Whilst lesser pointing restrictions may exist in the transfer phase, eclipse periods will differ from the operational case.

25.2 Operational Orbit Design

As introduced in [Sect. 4.4](#), although many types of operational orbit exist, certain types are prevalent; these are orbits close to Earth, generally used either for observing Earth or for communications, and geostationary orbits, whose primary use is for communications and meteorological satellites. A further interesting category is orbits around the Lagrange gravitational libration points of the Earth-Sun system for telescopes and solar observatories; further details can be found in [Sect. 4.4](#) and [1, 2]. This section describes some of the aspects of designing an operational orbit.

25.2.1 Low Earth Orbit

Many missions in low Earth orbit (LEO) are designed to observe aspects of the Earth. These could include global

surface coverage requirements, with the objective of overflying the same area within a designated period, or focusing on specific areas of the Earth (for example particular latitudes). Whilst many missions employ a single spacecraft, others use constellations. Constellations offer the advantage of decreasing the mean time between repeat observations of specific areas of the Earth (useful for disaster monitoring), or for telecommunications where global coverage is required at all times and when data relay between satellites can sometimes be used. A more detailed discussion of constellation design aspects can be found in [3].

As discussed in Sect. 4.4, missions in LEO are often designed to exploit certain key perturbations. This is particularly true for scientific, Earth observing missions. The best-known example is to maintain a near-fixed longitude difference between the Sun and the node of the satellite's orbit. These Sun-synchronous orbits allow the preservation of local solar time along the ground track of the satellite (the locus of the subsatellite point) and thus provide repetition of satellite illumination conditions from orbit to orbit. These perturbations and their effects are discussed in detail in Chap. 4, and for convenience are reviewed here.

25.2.1.1 Perturbations Due to Earth's Gravity Field

The Earth's mass distribution, such as that resulting from shape deviations from spherical and local density variations, results in deviations of the gravitational force from inverse square. The primary perturbation results from the fact that Earth is an oblate spheroid and as such possesses a significant J_2 gravitational harmonic. This has two main secular effects on an orbit: a rotation of the right ascension of the ascending node and also of the argument of pericenter

$$\dot{\Omega} = -\frac{3603}{\pi} \frac{3}{4} J_2 \left(\frac{r_E}{a}\right)^2 \sqrt{\frac{\mu}{a^3}} \frac{1}{(1-e^2)^2} \cos i \quad (25.1)$$

$$\dot{\omega} = \frac{1803}{\pi} \frac{3}{4} J_2 \left(\frac{r_E}{a}\right)^2 \sqrt{\frac{\mu}{a^3}} \frac{1}{(1-e^2)^2} (5 \cos^2 i - 1). \quad (25.2)$$

Both the ascending node and the argument of pericenter rotate at a steady rate which is strongly dependent on the inclination. An inclination can be found, using Eq. 4.141, where the pericenter does not rotate: inclination = 63.4°. This can be useful for specific latitude coverage in eccentric orbits (e.g. maintaining apocenter at a high latitude). As indicated in Eq. 4.139, a Sun-synchronous orbit can be achieved by modifying the inclination to achieve an ascending node rotation rate of 360° per year or approximately 1°/day.

The orbital elements all show a significant, regular periodic variation arising from J_2 . This is in addition to the long term (or secular) effects already described. Therefore,

a distinction is made between osculating orbit elements and mean elements

- *Osculating*—the instantaneous value at any point in an orbit. Used to define precisely the location of the spacecraft at a particular point in time.
- *Mean*—an averaged value over the orbit. Used in calculation or orbit periods and secular effects (see Chap. 4 for more details).

25.2.1.2 Perturbations Due to Atmospheric Drag

Atmospheric drag has a secular, perturbing effect on the spacecraft. It causes a steady reduction in orbital speed and hence semi-major axis. The drag force can be calculated by

$$\mathbf{D} = \frac{1}{2} A_p C_D V_{\text{atmosrel}} V_{\text{atmosrel}} \rho(h) \quad (25.3)$$

where C_D is the drag coefficient, ρ is the atmospheric density, $\mathbf{V}_{\text{atmosrel}}$ is the atmosphere relative velocity vector (i.e. the model assumes that nominally the atmosphere rotates with the Earth and any local wind speed effects can be included), and A_p is that area normal to the atmosphere relative velocity vector. Note that this equation is the same as Eq. 4.105, but is in a slightly different format. As discussed in Sect. 4.3.3, C_D typically takes a value in the range 2–2.5, although in certain situations higher values can be found. If drag is approximated as a constant force over the orbit then the change in semi-major axis, Δa , per orbit period, τ , is

$$\frac{\Delta a}{\tau} = 2 \sqrt{\frac{a^3}{\mu}} \sqrt{(1-e^2)} \frac{D}{m} \quad (25.4)$$

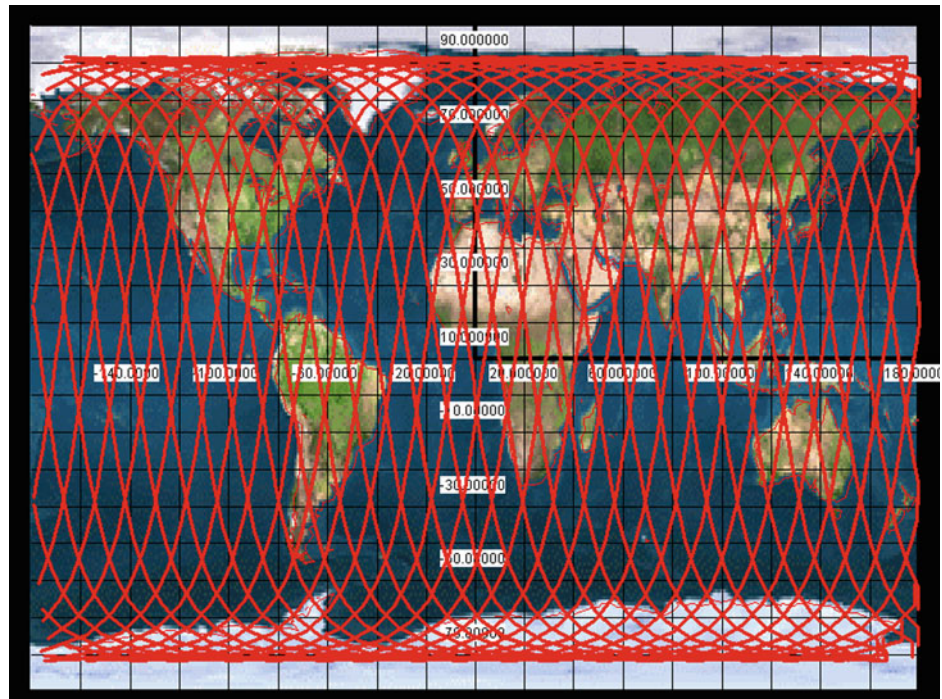
where eccentricity e is assumed to be $\ll 1$.

25.2.1.3 Mission Design in LEO

Missions in LEO are generally required to be Sun-synchronous, that is, the local solar time at the satellite equator crossing remains constant. The ascending node therefore rotates at approximately 1°/day. This can be achieved freely by exploiting the J_2 perturbation and choosing an inclination that yields the required nodal drift rate; see Eq. 4.139. The value is dependent on the semi-major axis of the orbit and ranges from typically 97° for altitudes at 400 km to 98° for altitudes at 800 km; see Fig. 4.13.

The orbit is generally required to be circular or near circular. In fact, so-called frozen orbits are often used to maintain an altitude profile that repeats with latitude (on both the ascending and descending node tracks). This can of course be achieved by a purely circular orbit but the presence of the J_2 perturbation means that the altitude will always vary throughout the course of the orbit, even if the eccentricity is instantaneously zero at some part of the orbit.

Fig. 25.1 Ground track for 727 km altitude circular orbit with a 2 day repeat period



The argument of pericenter is affected by J_2 , as described previously and also by the J_3 harmonic. However, it is possible to find an argument of pericenter and eccentricity that results in the frozen orbit condition. The mean argument of pericenter can be set to 90° and a mean eccentricity can then be found that results in no drift in the argument of pericenter (if predictions from only J_2 and J_3 effects are considered, see [4]).

In practice, other gravitational perturbations cause the argument of pericenter to exhibit a small periodic drift. The value for mean eccentricity required for the frozen condition is small, close to 0.0013.

LEO missions are often designed to have repeating ground tracks (subsattellite point on the Earth's surface). That is, after 'N' days the ground track overflies the initial ground track. This is achieved by selecting an appropriate orbit altitude, and therefore an orbit period. The longitude change in the ascending node at each successive equator crossing is determined by the orbit period (80–90 min for LEO) and the speed of the Earth's rotation. Values are typically $22\text{--}25^\circ$. It is important to note that longitude here is measured relative to the Greenwich (reference) meridian fixed on the rotating Earth.

For repetition to be achieved: $(M \times \text{change in longitude per orbit}) = (N \times 360^\circ)$. N is the number of days in the repeat period, the Earth rotating by 360° beneath the satellite in one day. Of course, M is an integer. Multiple altitude solutions generally exist for a given N.

- For $N = 2$, an altitude of 570 km gives a change in longitude of 24° and the ground track repeats after 30 orbits or 2 days.
- For $N = 2$, an altitude of 727 km gives a change in longitude of 24.27° and the ground track repeats after 29 orbits or 2 days.
- Further solutions exist for 28, 27, 26, etc. orbits.

Here altitude is the value of the mean semi-major axis minus Earth's equatorial radius (6,378 km). An example of a ground track repeating over 2 days is shown in Fig. 25.1.

25.2.1.4 Maintaining Low Earth Orbits

The previous section describing perturbations in LEO showed that atmospheric drag can slowly degrade an orbit, causing a steady reduction in altitude. This can be prevented by the application of regular, prograde orbit maintenance maneuvers. Typical intervals between such maneuvers can vary between a few days to several months. The interval is driven by the altitude (and hence the atmospheric density), and by the requirement to limit the change in altitude between maneuvers. Large changes cause a drift in the ground track relative to the reference case derived from an orbit that is unperturbed by drag, as the orbit progressively reduces in altitude. This means that overflight of the same latitude/longitude no longer occurs. A typical tolerance in the longitudinal drift in the ground track is 25 km.

The change of velocity, ΔV , requirement to maintain altitude depends on two main factors: altitude (and hence atmospheric density) and the area/mass of the satellite. Here

area is that projected into the direction of the velocity relative to the atmosphere and hence susceptible to drag. The ΔV per year can range between typically 100 m/s at lower altitudes (circa 350–400 km) down to less than 10 m/s (circa 800 km). The Sun's radiation in a specific wavelength range (around 10.7 cm) effects the Earth's atmosphere, and specifically the mean density with respect to altitude. The intensity of this emission varies over the Sun's eleven year cycle. The ΔV requirements for a specific mission depend strongly on the phase of the Sun's eleven year cycle relative to the mission start date.

A further orbit control maneuver can be needed to prevent the node of the orbit from drifting away from the value required to achieve Sun-synchronicity. Such a drift would result in a change in local solar time at the node. Often this is required to be controlled to within several minutes. The cause of this drift is a small change in inclination from the intended value, primarily as a result of lunar-solar gravity perturbations. A small out-of-plane maneuver can be applied. The maneuver frequency depends on the tolerance to drift in the local solar time, but typical values may be 1 maneuver per year, with a ΔV of 5 m/s. Discussions on orbit maintenance strategies for LEO missions can be found in [5].

25.2.2 Geostationary Satellites

As defined in Sect. 4.4.3, a geostationary Earth orbit (GEO) is one where the spacecraft appears fixed over a point on the equator, allowing continuous communications/observation as required. This is because the satellite travels around the Earth with the same orbital period as the Earth takes to spin on its axis.

A further, related category is the geosynchronous orbit. There are orbits with geostationary period but non-zero inclination. The resulting latitude/longitude motion of the subsatellite point depicts a lemniscate curve as shown in Fig. 4.9. This type of orbit can offer the possibility of extended range of latitude viewing.

25.2.2.1 Reaching Geostationary Orbit

Transfer to Geostationary orbit is generally achieved by use of intermediate orbits with maneuvers provided by the satellite, although some launch vehicles have the capability to directly inject into geostationary orbit. The most common strategy is however for the launch vehicle to inject the satellite into a geostationary transfer orbit (GTO), whose apocenter lies at geostationary altitude (35,786 km) and pericenter at typically 200–500 km. The satellite must then perform one or more pericenter raising maneuvers (similar

to the second half of the Hohmann transfer between circular orbits; see Chap. 4 for further details).

In some cases the launcher injection orbit does not have equatorial (i.e. zero) inclination but it is possible to combine the pericenter raising maneuvers with plane changing maneuvers if the argument of pericenter of the injection orbit is close to 180° . This means that the ΔV penalty for the plane change is relatively small compared to the ΔV needed for pericenter raising (which is typically 1,450 m/s).

25.2.2.2 Maintaining Geostationary Orbits

Once on its station (i.e. the target longitude), a geostationary satellite experiences perturbations: An east–west drift is experienced arising from the Earth's J_{22} harmonic (triaxiality). In addition, a north–south drift arises due to lunar/solar gravity effects. See Chap. 4 for more details on orbit perturbations.

The objective is generally to keep the satellite in a controlled 'box' of typically 0.1° in both latitude and longitude. Regular maneuvers are required to ensure that this is achieved. Typically, east–west stationkeeping maneuvers are performed at approximately fortnightly intervals, with north–south maneuvers being required less frequently. Some satellites are not required to maintain a north–south control and their inclination vector (i.e. inclination and right ascension of ascending node) drifts in a predetermined way over the course of the mission. Discussions on orbit maintenance strategies for GEO missions can be found in [5].

25.2.3 Interplanetary Missions

Interplanetary missions are very varied in their requirements. A typical format is the placing of a probe into an observation orbit around a planet or one of its moons. In some cases a probe is landed on the planet's surface. However, achieving the transfer between Earth and the planet in question is a key aspect of the mission design and often drives many aspects of the overall spacecraft design.

25.2.3.1 Basic Transfer Design

The simplest transfer between planets can be achieved via the solution of Lambert's problem, described in Sect. 4.5. The basic problem is the following

1. Leave planet A at a specified epoch (so defining the initial position vector).
2. Arrive at planet B at a specified epoch (defining the final position vector).
3. Solve Lambert's problem to determine the transfer orbit and therefore the velocities at planets A and B.

4. Calculate the spacecraft velocities relative to planets A and B and thereby derive the change in velocity needed to depart from planet A and to rendezvous with planet B.

Given a requirement for a velocity relative to a planet (on departing or arriving) an impulsive ΔV for the spacecraft maneuver may be derived. This is the ΔV to transfer from, for example, a defined initial bound orbit around the planet to a hyperbolic orbit, whose excess hyperbolic speed is equal to the magnitude of the required velocity relative to the planet. That is

$$\Delta V = \sqrt{\left(\frac{2\mu}{r_{pl1}} + V_{\infty}^2\right)} - \sqrt{2\mu\left(\frac{1}{r_{pl1}} - \frac{1}{(r_{pl1} + r_{apl1})}\right)} \quad (25.5)$$

for a general initial elliptical orbit with apocenter, a_{pl1} , pericenter r_{pl1} , and excess hyperbolic speed V . The constant μ is the planet's gravitational parameter. On leaving the planet's gravitational influence the residual speed relative to the planet is given by the excess hyperbolic speed. This is approximately equal to the magnitude of the required change in heliocentric velocity needed to instigate the transfer to planet B. For a further discussion see [2]. A similar maneuver is required for capture to a defined elliptical orbit about the target planet.

The epochs at which departure and arrival take place can be optimized to minimize the total ΔV (i.e. for escape plus capture maneuvers). For an idealized case, where the two planets were in circular, coplanar orbits, the epochs would be such that a Hohmann transfer can be executed.

25.2.3.2 General Interplanetary Transfers

In many cases, more complex routes are used to transfer between planets in order to reduce the total ΔV required. These can involve multiple gravity-assists at intermediate planets, as described in Sect. 4.5.6, and deep space maneuvers, in order to maximize the mass that can be injected into orbit around the target planet. Detailed discussion can be found in [2].

25.3 Concurrent Design of Space Missions

Any product or system comes into operation after a development process. This starts with the product's initial conception and proceeds with design, prototyping, testing, manufacturing and marketing (when required). The actual usage then follows and at the end of its life the product is disposed of. These phases are termed the life cycle. The development steps may be carried out sequentially, or there may be a degree of parallelism and integration.

Concurrent engineering is the discipline that looks at the techniques and processes that allow parallel execution of the product development steps. However, the term concurrent engineering is used with slightly different meanings within different engineering domains. For instance, the European Cooperation for Space Standardisation (ECSS) in [6] defines it as

Engineering activity taking place in the context of simultaneous design of the product, the production process and all associated product usages, in an integrated, multifunctional team, with external organizational constraints minimized.

Referring to the set of activities performed for selection, procurement, and management of the electrical, electronic and electromechanical (EEE) parts within a space project, the INCOSE (International Council of System Engineering) Glossary of Terms [7] provides a definition with wider scope

Engineering design practice that combines the concerns of marketing, functional product and process design, production, field service, recycling, and disposal into one integrated procedure.

In any case, all definitions maintain the two distinctive principles of (1) simultaneous execution of tasks, and (2) integration of team, tools, and processes. With this background, the concurrent engineering principles can be applied specifically to the design phase. Hence, the main characteristic of concurrent design is the quasi-simultaneous execution of the multiple tasks associated to the different engineering disciplines that compose the system design process. The second characteristic is the integrated character of the design process, i.e. the fact that the different design tools relevant to the different engineering disciplines are somehow linked together to form a coherent and consistent set (e.g. including same variables, units, conventions, margins, etc.). Finally, concurrent design emphasizes the involvement (and integration) in the design process of strictly non-technical disciplines taking into account from the start such external (to the technical domain) elements as cost and scheduling.

25.3.1 Historical Background

Concurrent engineering originated in the late nineteen eighties in the USA, as a new tool to shorten the development time of very large defense projects. The concept became popular worldwide in aeronautics and in the automotive industry. In Japanese carmakers for instance, concurrent engineering is presently commonplace and it is widely considered as one of the main contributors to their

success [8]. Concurrent engineering, and in particular concurrent design, for space applications was pioneered in the 1990s in the USA by NASA-JPL (Jet Propulsion Laboratory) [9] and the Aerospace Corporation [4] with the somewhat narrower scope of performing quickly feasibility assessments (i.e. pre-phase A) of space missions in order to identify the criticalities and cost drivers for later design and development phases. In Europe, the technique has been in use since 1997, at ESA/ESTEC (the European Space Agency Technology Centre) [10] from where it has spread to most large European space companies, universities and national agencies.

Although the principles have largely remained the same since the beginning, each organization has tailored the implementation of the methodology to its own needs and structure. For instance, some companies have adopted concurrent design as a tool for rapid design iterations, in response to invitations to tender. Agencies, on the contrary, tend to use it to make technical and economic feasibility assessments of space mission proposals coming, for instance, from the scientific community, and to help to define detailed mission and system level requirements for industrial design activities. Overall, there is a trend to apply concurrent engineering beyond feasibility studies into mission phases A and B, to enlarge the scope to different activities as design reviews and to cover targets other than mission design, as for instance, launcher or space instrument design, operations, etc. An example of successful application of concurrent engineering techniques to later phases of a space mission is the NASA Mars Pathfinder mission [11].

Recognizing the fostering role played by the ESA/ESTEC Concurrent Design Facility (CDF) in Europe, the following description takes the CDF as reference to explain the practical implementation of concurrent engineering techniques [12].

25.3.2 Space Mission Design

A space mission always comes in response to some expressed user needs, be they scientific, commercial, or institutional. The term ‘mission’ is used to indicate the overall set of tasks, duties, and functions necessary to accomplish the given user needs. ‘System’ is defined as the collection of functional elements organized together to perform the mission, i.e. the practical means used to perform the mission tasks. Within the system, and as discussed in Sect. 2.1.1, the ‘payload’ is distinguished from the ‘spacecraft’ (also called platform or bus). The payload is the part within the system whose operations allows the fulfillment of the mission tasks while spacecraft is here used to mean the collection of system parts that provide the

required services and resources for the payload to operate. Finally, the formal statement of a user need is a ‘requirement’. Requirements use invariably the ‘shall’ verbal form. It is common for the user to express also ‘constraints’. These are not strictly requirements but rather are restrictions in the possible system design solution due to financial considerations, risk mitigation, or political inputs. For instance, a common constraint in ESA space mission design is to limit the use of equipment subject to US procurement restrictions.

The space mission design process includes a large number of steps and many engineering disciplines, each dependent on the others for inputs and each providing outputs to be used by the others. The first step always consists of the definition of the mission requirements (i.e. which tasks shall be performed and how) and, as a follow-up, of the system requirements (i.e. which characteristics/performance are needed to execute the tasks). The definition of requirements involves interaction with the user/customer in order to translate the needs into technical language usable in the design process. In the early times of mission design, requirements definition is iterative, and is updated based on feedback from the system definition itself. In fact, the purpose of the initial design iterations is mostly to define the feasibility boundaries of the user’s needs and to update the mission and system requirements accordingly.

After a given set of mission requirements has been defined, the next step is generally the definition of the mission sequence to fulfill these requirements, from launch to operational orbit and ultimately spacecraft disposal. This is centered on the mission analysis discipline, which establishes all the orbital parameters and required maneuvers. Orbit(s) definition is the fundamental input in all the subsequent system analyses including

- Definition of mission environment (thermal, radiation, solar illumination, etc.).
- Definition of the reference launch vehicle.
- Definition of spacecraft’s nominal attitude(s).
- Definition of communications architecture (ground station selection, frequency plan, etc.).
- Definition of the systems modes of operation, for instance nominal mode with instruments operational, safe mode in case of contingency, eclipse mode in case of low power availability, etc.

The above tasks are normally carried out as a set of system-level trade-offs, where multiple options are compared based on parameters such as spacecraft mass, mission cost and risk, overall performance, etc. Once the main system-level parameters are available, the first iteration on the spacecraft design at subsystem level can start. This involves the execution in a predefined sequence of analyses involving the typical satellite design disciplines, as for instance,

attitude and orbit control, structures, thermal control, power system, avionics, ground operations, and so on.

During the first iteration, the design team defines the spacecraft's configuration (i.e. general shape and which equipment goes where) and initializes the system budget (mass, power, data, etc.). The first iteration generally results in the identification of several design issues with some mission and system requirements still in violation. Further design iterations address these issues and enter into more detail those areas that are considered critical or immature. If the process is managed properly, each iteration reduces the design issues and leads to a closer compliance with the requirements.

To assess convergence, the design team uses a few benchmark parameters; for instance, compliance to the requirements, total spacecraft mass in comparison to launcher performance, and the projected overall cost. If this set of parameters is deemed to have acceptable values, the design is declared completed. If convergence cannot be achieved, the initial mission and system requirements need to be renegotiated and some descoping is often agreed upon.

The process described above implies some degree of definition of the payload. Payload design is a discipline on its own, which normally progresses in parallel to the mission design and was introduced in [Chap. 6](#). A different team traditionally carries it out, as it requires different expertise from that employed for mission and spacecraft design. In this latter case, the interfaces between payload and spacecraft need to be defined. Those include a list of resources that the spacecraft is required to provide to the payload and a list of electrical, mechanical, and thermal characteristics to be maintained at the physical boundary between the payload and the spacecraft. These lists are controlled and, if necessary, updated after each iteration of the two parallel designs.

25.3.3 Concurrent Design

In the traditional mission design process, a design flow (i.e. which design task needs to be executed when) would be defined up-front by system engineering, who then request and control its execution in a sequential way by the different discipline specialists. These latter would generally work on their own and have knowledge only of the system design features of relevance to their work. The correct exchange of updated and consistent information between disciplines would rely solely on system engineering. In addition, the different discipline design tools are generally independent of each other, and outputs from one tool often need to be 'translated' before they can be used as inputs by another tool.

As mentioned, space mission design is by nature iterative; therefore this design sequence needs to be executed several times and convergence is often slow and painstaking.

Concurrent design replaces the sequential design with a simultaneous, i.e. parallel, process and requires all the design activities to be performed by a team that is physically or virtually co-located. Design takes place by all parties at the same location, at the same time. This has the following main advantages

- The role of information conveyor of the system engineer is marginalized by allowing the specialists to have direct interaction with each other and be aware of each other's design in real time. The system engineer can then concentrate on the nobler tasks of technical coordination and solution of system-level issues.
- Direct and immediate clarifications of design issues.
- Natural adoption of common approaches and rules among disciplines.
- More involvement and awareness of system level issues by the discipline specialists stimulating more creative solutions.
- Use of model-based system engineering techniques.

The tangible results of application of concurrent design are more rapid convergence and higher consistency, with a consequent reduction in the overall cost of the design process.

25.3.4 Implementation

Although the principles of simultaneity and integration typical of concurrent design are straightforward, their application is not, as it requires the availability and correct use of at least three main resources

- A system model.
- A multidisciplinary team working according to a well-specified set of rules and trained in the use of the model (management process).
- A formal process to carry out the activities (design process).

The design process is model based, meaning that the system model is used constantly as a reference for the updated description of the status of the design.

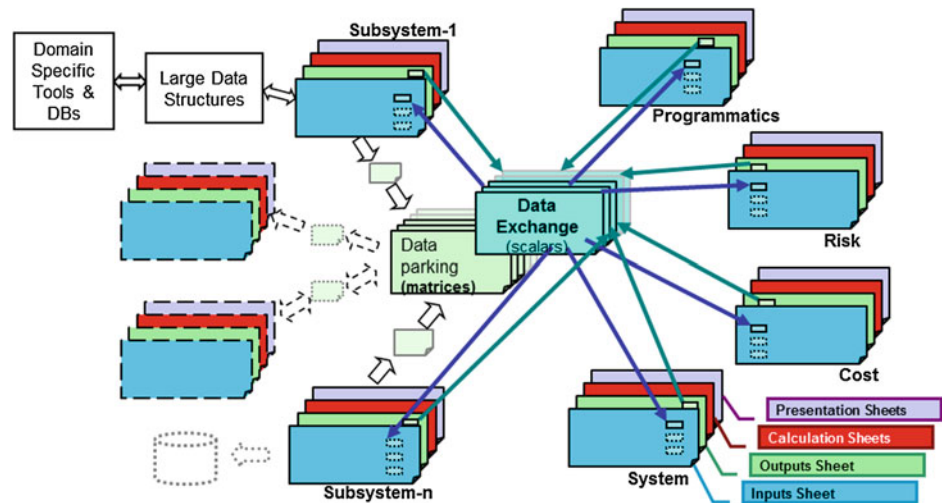
25.3.4.1 System Model

The system model is composed of three main elements

- An engineering database.
- A set of design tools.
- A software interface to allow access to the database from the different design tools.

The engineering database is a mathematical representation of the mission/system to be designed, in the form of

Fig. 25.2 Example of a system model: the ESTEC CDF integrated design model



scalars, vectors and matrixes representing all the different variables that characterize the mission/system. These variables include for instance

- The physical characteristics of system elements (units and associated masses, power consumptions, dimensions, etc.).
- Mission environment parameters such as orbital parameters, thermal and mechanical loads, radiation fluxes, etc.
- System global characteristics (total mass at launch, spacecraft dimensions, etc.).
- Requirements.
- Performance of the system.
- Any other information which is needed to assess cost or schedule as for instance the technology maturity level of the components of the system.

The variables shall be standardized in terms of definitions, names, units, reference coordinates, etc. The database provides the sharing of the data required for implementation of concurrent engineering principles. In addition, as it provides a real-time picture of the system, it is used as a tool to control the status of the design and to drive subsequent iterations. Further advantages are

- Possibility of extracting quickly system characteristics (budgets, configuration parameters, etc.).
- Possibility of performing parametric and *what-if* analyses.
- Possibility to create system performance simulations.
- Understanding the relationships between performance and requirements/constraints.

Finally, it is worth mentioning that the database provides the possibility to store (and to document, if the proper interfaces are added) the final results of the design activities in a coherent and uniform manner. This can be used in later design phases or as initialization for further mission studies.

The design tools may be in-house models or commercial software, and are used to size the different subsystems. Each organization has different preferences for its sizing tools, so the model must be flexible enough to allow different

software modules to be included. An important feature of the design tools is that they must allow for on-line design. Ideally, they should include different levels of complexity and model sophistication and rapidly provide first order results in order to permit the other disciplines to move in parallel.

As an example of a system model, the ESTEC CDF IDM (Integrated Design Model) is presented in Fig. 25.2. The database is represented by the light green and blue boxes (here called Data Exchange and Data Parking) in the middle, while the design tools for each discipline are shown all around; the design tools are represented by workbooks composed of a series of worksheets. The latter in turn may be connected to domain specific tools and other databases (e.g. commercial software for structural or thermal sizing, CAD tools, equipment supplier database, etc.).

Within each workbook, there is an area reserved for inputs and one reserved for outputs. Both areas are connected to the central database (green and purple lines). During the design, the team either retrieves required inputs or provides variables according to the responsibilities assigned to each specialist. In order to do so, a software interface between the design tools and the database needs to be implemented. This must include a mechanism for control of the data flow from and to the database in order to avoid continuous updates from the different disciplines that would result in uncontrolled design iterations. The process is therefore not automatic; to the contrary, human interaction and control is fundamental to ensure smooth progress.

25.3.4.2 Management Process

Use of the system model for design activities requires a well-defined process, namely a set of management and technical rules that must be followed by the team during the different design steps. Management rules concerns mostly

- The organization and the role of each component of the team, including decision-making rules (which are discussed below).

Fig. 25.3 Example of a concurrent design session at ESTEC CDF



- The logistics.
- The standardization of the output of the design activity, either in the form of data or documentation (reports, presentations, etc.). To exploit concurrent engineering practices, the design activity is normally organized in design sessions, e.g. plenary meetings with all team participants where the design occurs on-line and the design issues are discussed and analyzed. Off-line design should be allowed only when the level of complexity of the sizing activity requires simulation runs that cannot be completed within a session. This is to avoid updates of the database that are not understood/shared by the rest of the team. An example of a design session is shown in Fig. 25.3. The figure also shows a typical layout of a concurrent design room with a series of workstation all linked together and where the specialists sit; the team leader who directs the flow of information and controls the sequence of activities to be carried during a session takes center stage. Display screens facilitate conveying messages or explaining technical issues. Room(s) for splinter meeting(s) among reduced groups are normally also available.

25.3.4.3 Design Process

From the technical viewpoint the fundamental issue is to control the design iterations in order to allow smooth design steps and fast convergence. As already mentioned, space mission design is highly iterative, requiring several system-level loops with many lower level loops nested inside. All these loops can be efficiently controlled by defining a set of rules for accessing the engineering database. Each successive version of the database is the result of a given predefined set of design activities and would constitute an iteration. At the end of each iteration, the team inspects and analyzes the content of the database and extracts system-level information such as total mass, cost, risks, performance, compliance to requirements in order to identify design issues and target the following iteration.

Another important element of the concurrent design process is the way technical trade-offs are handled. In concurrent engineering, the trade-offs and the associated options and criteria are defined by the team all together. Hence, decisions are shared and transparent, and the justification for design choices are more transparent. In addition, due to the model-based approach trade-off outcomes are well documented and archived for easy retrieval at later stages of the project.

Concurrent design has proven very effective also in the definition and management of the margins. In space mission design, margins at different levels are applied to take into account uncertainties such as an ill-defined environment (thermal, mechanical, radiation, etc.); the poor fidelity of analytical models, unforeseen factors, simplifications, etc.; at the equipment level coping with technology immaturity (i.e. if a unit is not qualified yet, there is an uncertainty on its actual performance); and at the system level coping with uncertainty in the requirements, in the definition of the payload interfaces, and in the launcher performance.

All these different margins get applied one on top of the other, and if not properly and consistently controlled they can make the design overly conservative or unreliably optimistic. Concurrent design, due to its integrated and model-based nature, allows for a better and more consistent definition and control of the margins.

25.3.5 The Team

Concurrent engineering emphasizes the role of a team, which assumes a more central role and responsibility compared to the traditional design approach. A few points are specific to concurrent engineering and will be discussed below. First, the concept on design on-line allows the direct participation of the customer/user to the design activities. This greatly accelerates decision-making, permits direct clarification of the requirements and immediate redirection in the event of misunderstandings. In addition, all technical

decisions are transparent and shared. A second point is the role of the team leader. This is a management figure distinct from the system engineer and whose function is to ‘direct’ the design session by making sure that all discussions and analyses remain in the scope. He/she acts as a moderator and has the authority to take final decisions in consultation with the user/customer. Thirdly, there is the inclusion in the team of ‘non-technical’ (strictly speaking) disciplines such as cost evaluation, risk assessment, and scheduling, which now have full visibility of the design process and can perform on-line assessments. This is in contrast to the traditional approach, where these evaluations take place a posteriori, i.e. only when the technical design is consolidated. This avoids the possibility of picking design choices that are efficient technically but unacceptable from the cost, risk, or schedule viewpoints.

Risk is taken into account within the design process by performing, in parallel to the design, a risk analysis, i.e. identifying those events whose occurrence is perceived likely to cause loss or degradation of the mission objectives. These events can be of technical or programmatic nature. For instance, the use in the design of a technology of low maturity presents a risk to the project because its development for flight may cause delays and additional costs. The risks are classified on the basis of their severity (how damaging their occurrence would be) as well as on the likelihood of their occurrence, assessed either on statistical data or on experience. The design team is then instructed to avoid design solutions that could cause the occurrence of a risk with a high severity and a high likelihood or, in case this is not possible, to provide mitigation strategies; e.g. in the case of an immature technology, select as an alternative a more mature but less performing technology.

Finally, specialists representing the later phases of system development such as verification/testing and operations complete the team. They work together with the design team from the beginning, taking care to pass lessons learned from other projects, and to define the initial test plan and operational concept.

25.4 Low Cost Mission Design

One of the limiting factors to involvement in space related activities has been the high cost associated with traditional missions. In the early years of space exploration, this was not a particular problem as the majority of missions were run by space agencies and the military, which were directly funded by their associated governments. Though involvement in exploitation space began to grow considerably in the 1970s, the costs involved still tended to favor governments and large commercial corporations. The goal of opening up space to more diverse organizations could only

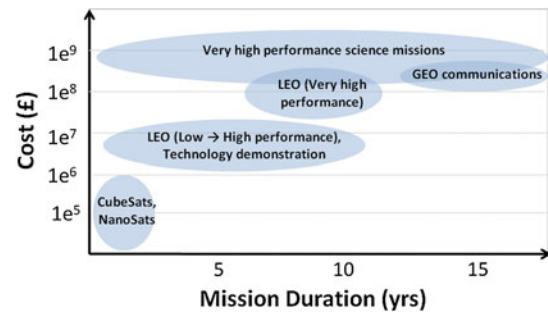


Fig. 25.4 Some examples of typical space applications and estimated costs

be achieved if lower cost alternatives were sought. The result today is that numerous missions have been created to stimulate alternative markets, mission types, and objectives by providing much lower cost access to space. Indeed, the desire for cost-effective missions across all areas of space technologies and mission types is even more important in today’s global financial environment. Commercial organizations need robust business plans to gain funding, and even government funded missions are seeking more cost-effective missions when their budgets are cut.

25.4.1 Low Cost Missions and the 80/20 Rule

Before considering how to design a low cost mission, it is important to categorize what a low cost mission means. There are numerous applications that use space nowadays, and they vary dramatically in their objectives and hence cost. When referring to the cost of the mission, it is the total cost required to fund the mission from its earliest concept design right through to in-orbit operations and eventual disposal. It includes all design, manufacturing, management, testing and operational aspects of the space and ground segments. Figure 25.4 illustrates some typical examples of mission types varying from very low cost CubeSat missions (costing thousands of dollars), typically used for educational or technology demonstration purposes, right up to high performance space agency-run science missions such as Cassini-Huygens. Crewed space missions are not illustrated because these are a further order of magnitude (or two) more expensive.

This section focuses on the design of those missions towards the lower end of the range of mission cost. That is to say, medium duration (<10 years) operational missions primarily for commercial or scientific purposes or even advanced technology demonstration or LEO communications. These missions are typically focused on (but not limited to) microsatellite applications with a mission cost of several million up to several tens of millions of dollars for

single spacecraft depending upon the mission objectives and spacecraft capability. The reasons for this definition of a low cost mission will become apparent during this section, but in summary the highest performance missions have the most rigid and demanding requirements, and they offer very limited opportunities to produce equivalent low cost missions. In addition, although the lowest cost missions (CubeSats) are not addressed within this chapter, it is widely held that in coming years technology developments will mean they become as capable as current microsatellites, thereby opening up even more possibilities for ultra-low cost missions. Furthermore, another distinction will be made. The term 'low cost' itself must be defined. Within the mission class defined previously, a low cost mission is one which is significantly cheaper ($\ll 50\%$) than a comparable mission developed in the traditional manner. This comparison is important, because the general design process for both options would look similar at a high level. Therefore, in order to illustrate the subtleties of a low cost mission design, a comparison will be made throughout with the traditional industry approach. In the context of this section, the two approaches can be briefly summarized as follows

- Traditional space mission: A mission typically designed to be fully compliant with a series of detailed space engineering standards (e.g. ECSS, MIL). They typically follow a very low risk design, build and manufacturing schedule, especially for new developments and bespoke missions. These types of mission are typically built for space agencies, military, governments, and large commercial operators.
- A low cost mission: A mission design characterized by a less process driven development and by more rapid timescales, with a different approach to risk management. There are typically smaller spacecraft for a broad customer base depending upon the mission objectives. Customers can include smaller commercial operators, developing space nations for operational missions (for example Earth observation), and sometimes even space agencies and the military (typically for educational and technology demonstration missions).

The distinction between the low cost and traditional approaches is important. The general set of steps used to design and construct a mission by either approach will look very similar in terms of requirements analysis, design, manufacture, test etc. As such, this section does not repeat this procedure for low cost missions, but rather highlights where differences between the two approaches can lead to a reduction in the total mission cost.

The question remains: How can you design low cost space missions? There is not a straightforward answer to this, and indeed there is no secret formula that will always significantly reduce the cost. In reality, it is more of a design philosophy, where the application of many concepts

together can lead to a significant reduction in mission cost for certain applications. Some of these concepts will be familiar to engineers, and will already be part of day-to-day operations of space manufacturing organizations. However, a flexible, pragmatic approach is needed because the many interrelated elements may require to be applied in subtly different ways for diverse sets of mission objectives. Only when these many smaller things are implemented together can space mission cost be dramatically reduced.

The fundamental element for the design of a low cost mission is the 80/20 rule. This states that for a given set of mission objectives, 80 % of the required performance can be achieved for 20 % of the mission cost. Therefore, if there is some flexibility in a customer's aims there is ample opportunity to vastly reduce the mission cost. To fully apply the rule, two important steps need to be taken: (1) identifying the type of customer and mission objectives and (2) challenging the requirements.

25.4.1.1 Customer and Mission Type

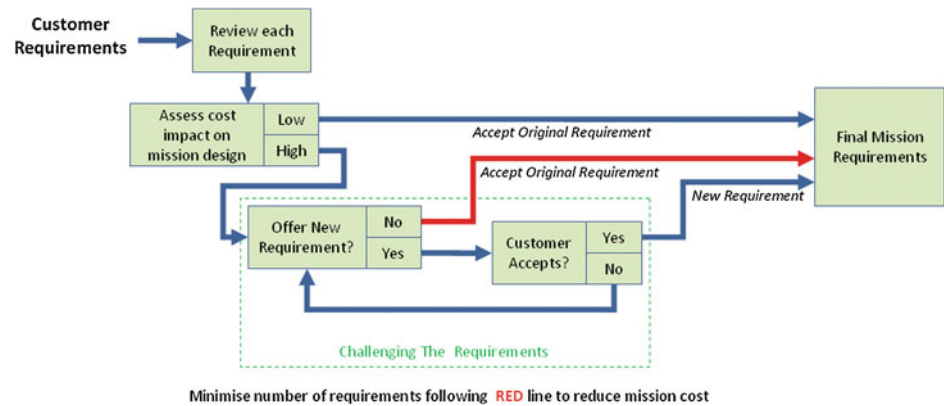
The customer and purpose of the mission is critical to understanding whether the 80/20 rule can be applied successfully. Take for example a space agency wishing to develop a high performance mission with many detailed scientific goals. Such a mission will typically have a comprehensive list of rigid requirements that may have taken several years to develop. These requirements will represent a balance between the aims and considerations of the customer (the space agency) and the end users (the science community). No doubt, the end aims of the mission will have been one of the key elements to ensure that it was selected for development in the first place. Such missions also generally follow the traditional minimum-risk approach, requiring adherence to numerous standards that define a multitude of additional requirements for the design, manufacture, and quality of the mission. All these things combine to limit the scope to negotiate or 'challenge the requirements' in an effort to apply the 80/20 rule. Therefore, the low cost mission needs to focus on other opportunities where there is more flexibility in the possible design. This could be, for example, through a less rigid set of mission objectives or a customer who is prepared to accept alternative development approaches. This leads directly to the idea of 'challenging the requirements'.

25.4.1.2 Challenging the Requirements

The design of any mission will start from a set of mission requirements or objectives of which the detail and quantity will vary considerably from mission to mission. However, it is important to remember that these requirements are not just limited to the technical elements.

- Technical specification: Defining the overall mission objectives, the main technical parameters of the

Fig. 25.5 Process of challenging the requirements for low cost mission design



spacecraft and ground segment and operational requirements. May also include requirements on how the system should be tested, verified, etc.

- Programmatic elements: Defining schedules, cost, documentation requirements, meetings, penalty clauses and quality.
- Other: There could be additional constraints depending on the customer. For example, political considerations such as the location of ground stations or requirements to provide training.

It is therefore important not only to focus on the technical aspects when challenging the requirements but also on programmatic elements if a truly low cost mission is to be developed. The process of challenging the requirements is simple in theory; identify the primary mission requirements or objectives that are essential for the customer and those that are secondary ‘nice-to-haves’. This will allow the mission designer to gain a good understanding of the available trade-space, and will indicate the areas of the mission design where there may be scope to relax or realign certain requirements or parameters. A detailed assessment can then be undertaken to identify which requirements actually drive the cost of the mission design. This could be through costly bought-in equipment, extensive new developments, or the use of a higher cost launch opportunity to name but a few examples. If some of these driving requirements relate to areas of secondary importance, then negotiations with the customer can commence on ways to create a reduction in mission cost. The result should be that the customer is satisfied with the final set of agreed requirements, even though they represent 80 % of the original performance envisaged. However, a new mission cost of only 20 % of the original prediction is clearly highly satisfactory and may actually be the difference between the mission receiving funding or not. This process can be seen in Fig. 25.5.

As stated, challenging the requirements should not just be limited to technical matters, but programmatic elements

as well. Consider for example the level of formal review points and associated documentation that is to be delivered to a customer. The more documentation that is required and the more frequent the formal project reviews are, then the less time is available for each engineer to actually perform their day-to-day engineering activities. This will mean that a larger team is needed in order to maintain the project milestones and the cost will increase. Therefore, the ultimate goal is to find the minimum level of documentation with which the customer can review the progress of the mission satisfactorily or operate the satellite without problem.

As an example, consider a hypothetical set of requirements for a new Earth observation mission. It is assumed that an existing spacecraft will be proposed for the mission and that any changes to the design will be identified. This existing spacecraft design is an Earth observation platform that currently operates at 690 km, giving a ground resolution for the imager of 2.5 m. Each requirement has to be assessed to see if the current spacecraft is compliant. If not, the impact on the mission design is assessed to determine if the requirement should be challenged.

The example in Table 25.1 illustrates the process of challenging the requirements for a few fictitious requirements. In reality, the process will require plenty of negotiations with the customer. Compromises will be needed between the designer and customer to find a solution that the customer will accept and which will not impose too many high risk developments on the manufacturer.

25.4.2 Development Approach

The low cost development approach needs to consider evolutionary development: (maximizing reuse of existing equipment and software), standards, component selection, and testing and qualification.

Table 25.1 Example of requirements challenging

Initial requirement	Impact	Offer new requirement?	New requirement to offer
Schedule: 24 months to launch	Medium: current predicted schedule of the existing spacecraft: 24 months to be ready to launch, provided any new developments can be maintained within this timescale and a launch can be found in time	Yes: there is always uncertainty when a launch will occur and it is generally out-with the manufacturers control	Schedule: 24 months to flight readiness review (this maintains the same schedule for the construction of the spacecraft but decouples it from the actual launch date)
Ground sample distance: 2.25 m	Medium: current imager has 2.5 m GSD from 690 km orbit	No: the current spacecraft can operate at ~620 km altitude which will result in a GSD matching the requirement so this acceptable	n/a
Maximum off-pointing: $\pm 35^\circ$ from Nadir	Low: current spacecraft can achieve this as standard	No: the proposed spacecraft is already compliant	n/a
Maximum image strip: 5,000 km	High: current spacecraft can achieve a maximum image length of 2,500 km	Yes: to achieve this will require a new design of the payload electronics to allow it to operate for this duration of time continuously without overheating and getting degraded performance. Additional on-board data capability storage is also be required	Maximum image strip: 2,500 km (If the customer accepts a shorter maximum imaging strip then the current spacecraft can remain unmodified and therefore will result in a lower cost and fewer developments)
De-orbit: at end of life the spacecraft must ensure atmospheric re-entry within 10 years	High: current spacecraft does not contain sufficient propellant for de-orbit in the required timescale. However, it should be able to de-orbit within 25 years	Yes: such a requirement would require an enhanced propulsion system or the inclusion of de-orbit device. The impact could be an accommodation issue (larger propellant tanks, or inclusion of a de-orbit device such as a tether) or may require the development of a more capable propulsion system	De-orbit within 25 years (without relaxation of the requirement, new developments or externally procured equipment may be needed. This will increase mission cost and risk)

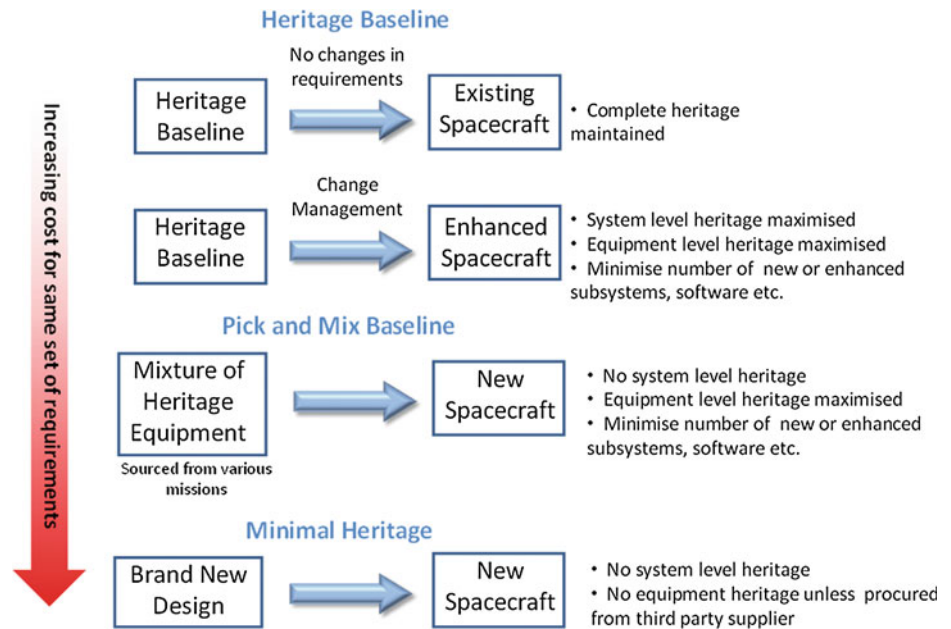
25.4.2.1 Evolutionary Development Approach

To aid with the design of a truly low cost mission, another important consideration is to perform an evolutionary development approach. This essentially means that the new mission, spacecraft, or subsystem is derived from a previous one, rather than starting with a brand new design. This relates directly to the earlier discussion of the 80/20 rule, whereby challenging the requirements can lead to more direct heritage. Therefore, if a customer can see how changes to some requirements allow more extensive heritage from a previous mission, then the overall cost and indeed the risk of the mission can be dramatically reduced. Again, starting a mission design from an existing spacecraft is nothing new, but in combination with challenging the requirements in order to achieve the 80/20 rule by minimizing new developments and requirements for higher performance equipment it can yield large cost savings. Note that this example just illustrates the process for the spacecraft design, but it can also be applied to the programmatic elements, ground segment, and/or operations.

In Fig. 25.6, it is not surprising that the lowest cost approach would be to produce an exact rebuild of a previous spacecraft. However, in reality this is very rare. There are usually some differences from mission to mission, which

may be caused by many things. These could include operation in a different environment, an alternative launch vehicle, or larger changes to requirements such as increased downlink rates or improved agility. As such, the existing spacecraft will usually require a certain level of enhancement and therefore the process of change management can be used to control these modifications without dramatically altering the way that the overall system is designed and operated. It is accepted that the approach of maximizing equipment heritage is very low risk at the equipment or even the subsystem level, but the cost impact comes when these individual elements are combined into a complete system. This system may not have heritage as a whole, and therefore additional cost can result through a large amount of testing in AIT. This testing will be vital to ensure that these elements can operate smoothly together. The final starting point of a completely new design is another area that will not typically occur. It might apply to a brand new startup organization or to an educational mission, but not to operational missions because the risk would be too high for established organizations. What these simple examples illustrate is that through the application of the 80/20 rule and challenging the requirements, it is possible to maximize

Fig. 25.6 Development approaches for new missions



direct heritage from a previous mission and therefore reduce risk and cost.

Deviations from this evolutionary development approach will begin to occur as the mission requirements move further away from an existing spacecraft design or mission. It may be that the new mission requires a payload far in excess of the physical size and mass previously flown on an existing spacecraft. In this case, there may be no option but to propose a brand new structural configuration for the spacecraft. However, even though mechanically the design might be a new development, maximum heritage can still be gained from using existing avionics, software, and even operational concepts.

25.4.2.2 Standards

Common standards have been developed over the years to achieve interoperability in order to ensure that products meet certain requirements for quality, commonality, and reliability. In space engineering, these standards cover all aspects of mission management, design, manufacturing, and operation. They have been produced to cope with the most complex and demanding missions which could be created. As a result, they contain comprehensive and rigorous requirements and procedures to provide a low risk approach to be pursued when developing these types of missions. The standards used vary globally, but two of the most widely used are the US Mil Standards and European ECSS. Although appropriate for the highly challenging missions for which they were designed, these standards do not necessarily downscale effectively or efficiently for smaller, low cost missions. Therefore, to attempt to apply such standards for

less demanding mission may generate a significant amount unnecessary work. This could be the result of a requirement for more extensive analysis and testing, which in turn results in higher costs and longer schedules. Even if only a subset of the approximately 120 ECSS engineering standards are used for an particular mission, it will take a vast effort to demonstrate compliance to the many requirements contained within each standard. A lower cost alternative is for the mission design team to identify the tests and procedures that are absolutely necessary (i.e. add value) within the scope of the mission without increasing risk to an unacceptable level. This can be done through the design of company specific procedures, or even taking applicable parts from existing standards. Indeed many lower level standards (e.g. ECSS-Q-ST-70-08C—*Manual soldering of high-reliability electrical connections*) are perfectly applicable and are used for either traditional or low cost mission design approaches as they relate to the quality of the manufactured parts.

The most important thing is to have sufficient flexibility to be able to determine the most suitable approach that allows the production and operation of equipment at a necessary level of quality to satisfy the mission goals. To understand how to do this in reality requires experience from the engineers within the mission team, and sufficient flexibility in their organization to adapt to positive changes. It is therefore not something that can be learned quickly. However, the key message to convey about the use of standards for low cost mission design is to think about what is necessary within the schedule and risk constraints of the mission. Question whether a particular document, test, or analysis adds value or reduces risk and does not simply

Table 25.2 Comparison between space qualified and COTS components

Component	Typical applications	Notes
Space qualified or high reliability	Space missions, military applications	<p><i>Pros:</i> They are certified. E.g. by ESA, MIL-S etc. High level of traceability to wafer level, fully tested, screened, those specifically designed for space use are usually radiation assured. Therefore, generally ready to use when purchased</p> <p><i>Cons:</i> Built in low volumes and a high procurement cost, longer lead times, Some US components may be ITAR controlled. The screening process can be hazardous</p>
COTS	Commercial, industrial and automotive industries	<p><i>Pros:</i> Inherently reliable due to tight process control. No lot based qualification. Built in high volumes so well proven and controlled processes. Most advanced and up-to-date technology available. Low cost and short lead times</p> <p><i>Cons:</i> No screening. Not full traceability to die. Far more susceptible to the space environment than other components</p>

increase mission cost. Note that using fewer standards for low cost missions does not result in a lower quality; it just produces a more streamlined approach.

25.4.3 Component Selection

Another extremely important area to address when designing a low cost mission is the choice of the type of components used to create the spacecraft avionics. These can be loosely grouped into two types of electrical components commonly used in space; namely space qualified or high reliability (as discussed in Sect. 3.4.4), and commercial-off-the-shelf (COTS).

The differences between these types of components is summarized in Table 25.2.

To design a low cost mission, COTS components are attractive because they are significantly cheaper than their space-grade equivalents. Despite not being designed to operate in space, they have been successfully used on many low cost space missions. This provides the best possible proof that they can be used in such an environment. A significant advantage with using COTS components is the flexibility afforded by short lead times and low cost. This means that components can be rapidly changed if they are damaged during testing or if a design modification requires an alternative part. If a comparable space-grade component was used with a lead time that could well be in excess of a year, then ordering a new component would not be possible without a major impact on a project schedule. It should also be noted that the use of COTS components will generally only result in a much lower equipment cost than using space qualified components provided that the mission can satisfy the following constraints

- Radiation total dose is low: COTS components will typically survive up to 5–10 Krads(Si) meaning they are best suited to short to medium duration LEO missions (<7–10 years) depending on the exact orbit (though this does not exclude their use beyond LEO in certain cases).

- The effects of single-event effects (SEE), as discussed in Sect. 3.3.2, on the components are understood and will not impact the desired performance of the system. For example, if a single event upset (SEU) causes an equipment reset, then the system will be robust enough to maintain safe operation while the equipment reboots.

Therefore, before using COTS components, the operational environment should be assessed carefully and the implications considered across the whole project. While the raw cost of the components is significantly lower for COTS compared to space-grade equivalents if the qualification approach used requires every COTS component to be extensively tested (e.g. for radiation ‘hardness’), then the overall cost to the mission could actually be higher than buying an equivalent qualified space-grade component. The detailed procedure for component selection is beyond the scope of this chapter but to summarize, consider the following criteria

1. Driving requirements (cost, schedule, performance).
2. Environmental assessment (radiation, thermal etc.).
3. System design implications (FDIR, redundancy, duty cycle).
4. Qualification approach (test every new component, flight heritage in comparable environment).

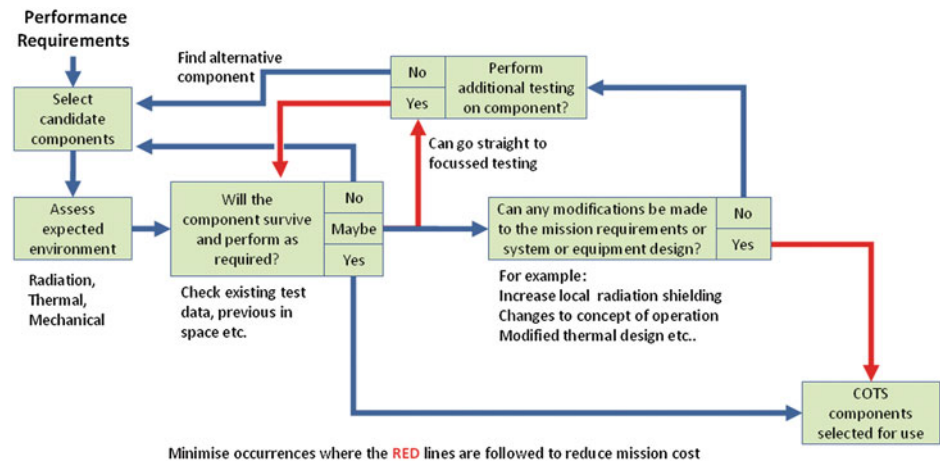
A simplified view of an example selection process can be seen in Fig. 25.7.

In summary, COTS components can provide the greatest flexibility, highest performance (through access to most up-to-date-technology) and lowest cost option for a space mission design so long as the mission design and test approach is suitably selected. For long duration, high radiation missions, it is probably cheaper to use space-grade components in many cases rather than performing extensive testing and qualification on COTS components.

25.4.4 Testing and Qualification

Once the heritage baseline has been identified and changes between this and the new mission identified and agreed,

Fig. 25.7 Example of COTS parts selection process for an avionics unit



then careful management of these changes is needed to ensure that the mission cost remains low. This management also has to consider the testing and hence the qualification approach used for the spacecraft.

The conventional approach in the space industry is to perform rigorous low level testing in order to qualify all equipment and subsystems to any new requirements, before advancing to high level testing of the full system. There is nothing wrong with this approach if a conservative approach to risk management is desired. It may even be dictated by contractual interfaces, where a subcontractor is required to 'prove' their equipment works before it leaves the factory. However, when designing a low cost mission this strategy can be addressed in a slightly different way. Rather than qualifying all equipment through low level testing, restrict this approach to just new developments, or equipment which has undergone substantial modification (of either hardware or software). This more focused testing is used to retire the largest risks as early as possible in the project, particularly before the spacecraft enters AIT where large modifications can be far more costly. For all other internally manufactured subsystems, these can be verified at a higher level, as part of the whole system. Note that for organizations that have a large amount of externally procured equipment, the suppliers will have qualified their units prior to delivery to AIT. This approach is illustrated in Fig. 25.8. Those that are heavily modified, or are new developments will typically have an engineering model (EM) made. This may simply be the electronics, and testing is performed to ensure that the design operates as expected. After this, if it is deemed too high risk to proceed directly to the equipment proto-flight model (PFM), then an engineering qualification model (EQM) can be manufactured. This is basically a flight representative module, electrically and mechanically, which will be tested thoroughly in order to qualify the unit for its expected environment. Indeed, the unit could even be tested at system level on the structural qualification model (SQM) where it would experience a

very similar mechanical environment to that expected on the actual spacecraft. After this point, a flight model (FM) of the equipment can be built to be integrated on the spacecraft. Alternatively, if the risk of mechanical failure was not deemed high, the unit could proceed directly to a PFM without the need (and cost) of an EQM. This unit would undergo some functional testing before integration onto the spacecraft. Indeed, all equipment/units undergo functional testing before being integrated into the spacecraft PFM. This is to ensure that it is all operating as required prior to integration.

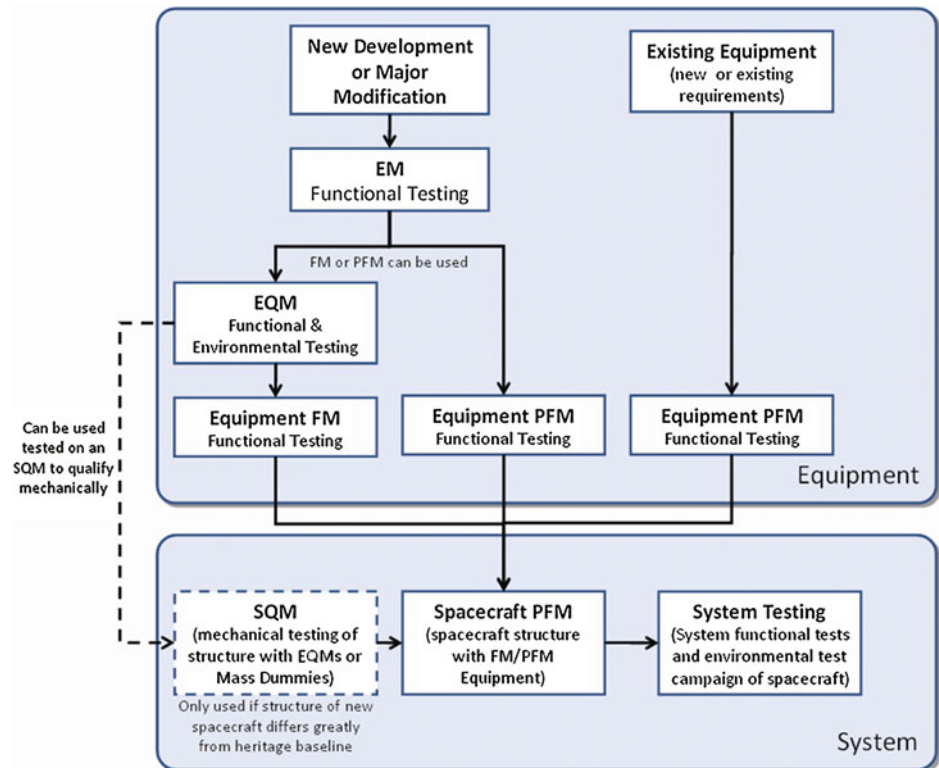
The main aim of the equipment level testing is to demonstrate that none of the equipment will damage the spacecraft when integrated into a complete system, as well as to verify those requirements that cannot occur at the system level. In contrast, the system level testing is used largely to verify the performance of the entire system including ground segment.

The advantage of this approach is that as much of the testing as possible is performed in the flight configuration. This means that the spacecraft and its software can be operated in as realistic a manner as possible prior to launch. All interfaces can be thoroughly exercised, ideally using actual hardware and software from the operational ground station. This allows commands to be generated and data to be returned from the spacecraft in exactly the same manner as if it were in orbit. Obviously not all equipment can be tested in such a representative manner (e.g. propulsion systems, AOCS equipment), but a large proportion and the spacecraft equipment and subsystems can be verified in this way.

25.4.5 Risk Analysis

Another significant way to reduce cost is by the manner that risk is handled within the project. This means that key risks specific to the mission are identified and effort is focused

Fig. 25.8 Example of test flow and model philosophy



only on addressing those. Effort is not spent reducing every imaginable risk to zero prior to launch, no matter how low the probability of it actually occurring. This does not mean that the mission is any less likely to fulfill its mission objectives than a traditional mission. All it means is that a less cautious approach can be adopted with regards to things like reliability, testing, qualification, etc.

It is important not to induce rising costs by excessive risk reduction for the type of mission. Not all risks can be completely retired during a mission, since there are so many unexpected events that can happen throughout the design, manufacture, and operational phases. In addition, no spacecraft can really be 100% reliable or offer 100% availability (though some come very close). The key is to find a suitable balance between an overly optimistic approach (high risk) and a very conservative approach (lowest possible risk). Examples can be found of both extremes producing undesirable results, ranging from tragic failures of human space flight missions, built to the highest safety standards and minimum risk, to failures at the other extreme such as exceptionally low cost student-built satellites. In addition, irrespective of the type of spacecraft, risk philosophy, etc., there is still a relatively high possibility of a launch failure. Indeed current launch systems are averaging a success rate of the order of 91%. With this in mind, risk can be managed in a more pragmatic way. The general approach to minimize costs is to focus efforts on tackling

the major risks and developing appropriate mitigation strategies. The way this is achieved will be described in more detail in the following sections.

25.4.6 Schedule Implications

It has already been mentioned that a low cost approach to risk revolves around focusing on the most critical risks rather than trying to eliminate all risks. This produces a more streamlined approach that makes the best use of the available resources without extending project schedules. To minimize mission cost it is important to keep the overall schedule as short as possible without increasing risk beyond an acceptable level. This is for several reasons. Firstly, many low cost missions rely on secondary launches. As such the launch date is always driven by the primary payload on the vehicle and therefore, if the other passenger is not ready it may miss its launch. In addition, the longer the schedule, the longer the team of engineers and managers will be working on the project, increasing cost. However, if the schedule is shorter than some optimal point, there may be little time to allow for additional testing. The result could be that late-breaking problems arise in AIT. These are typically very costly to fix that late in the project and again will drive up mission costs considerably. Therefore, by gaining an early understanding of the critical risks and appropriate mitigation strategies, it should be possible to

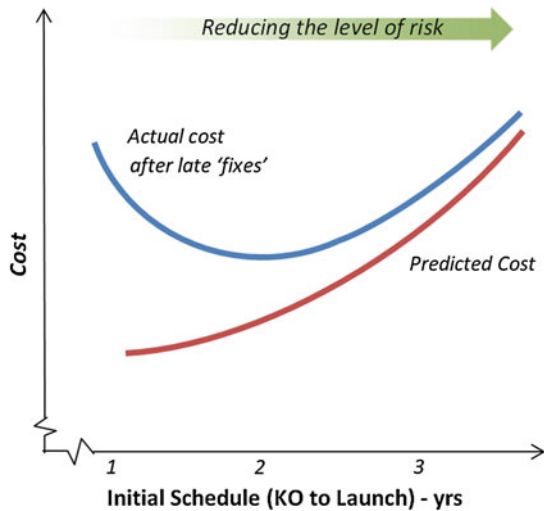


Fig. 25.9 Cost and risk for an arbitrary mission if different schedules are used

determine the optimal schedule for the mission. An example of this general profile is seen in Fig. 25.9 where the initial cost of the mission is shown for a variety of possible schedules. To find the overall lowest cost mission, the balance between schedule and risk can be seen. A minimal schedule might lead to late breaking issues towards the end of manufacture and test (too many corners cut), which will increase the overall actual cost. Similarly, a conservative schedule due to rigorous testing throughout to vastly reduce risk will also be costly, but with a far lower likelihood of late breaking issues. The key is to find the optimal balance between the two extremes.

The profile in Fig. 25.9 will vary considerably from mission to mission because there are always so many unpredictable events throughout the mission development, and even the most experienced teams pursuing the most conservative approach to risk management can experience costly late breaking issues.

25.4.7 Risk Management

The next area to address is an approach to actually managing risk whereby the project will only focus on those deemed most critical. The standard approach in the space industry is to create a risk register; see Sect. 7.8. This requires all engineers and managers on the project to identify any risks that they foresee and to rate them according to the likelihood of their occurrence and the severity of their impact. In addition, suggested mitigation strategies are needed of their impact into the revised estimates of cost and schedule. Three example risks are shown in Table 25.3, though in reality there could be hundreds identified on any particular project; that is why it is so

important to focus only on those deemed to be most critical if mission costs are to be minimized.

To focus the main effort on the critical risks requires a ‘bottom-up’ approach. This revolves around managing risks at the lowest level in the project hierarchy. Therefore, those people involved with a particular work package will identify all the perceived risks to their work. Those that are deemed critical (high probability and large impact to the project) will be flowed up to the project level risk register, to be managed by the project manager. It may even be that the project manager flows a further subset of risks to a corporate level to be dealt with by company directors if the risk is severe enough or cannot be managed or mitigated at project level. All risks are reassessed at regular review points and can be retired as the mission progresses. Of course, as risks are retired, new ones can arise, so regular risk reviews are needed in order to ensure resources on the project are correctly focused on the right areas (Fig. 25.10).

It is important to stress that accepting an alternative risk management approach in order to reduce mission cost, does not mean that it is more likely to fail than a traditional mission. The low cost mission simply relies on focusing the design effort on the most critical areas to save time and resources (hence money).

25.4.8 Margins and Design Flexibility

Consider now some of the engineering approaches that are used during the design phase to manage risks that may appear later in the project: (1) application of suitable margins and (2) flexibility in design.

25.4.8.1 Margin Philosophy

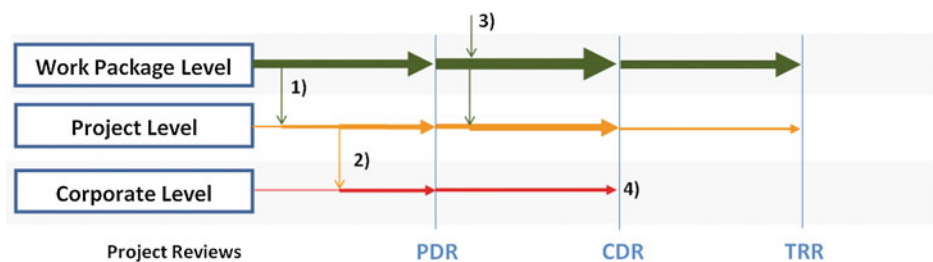
The application of suitable margins for mass, power consumption, data storage etc. is of course standard engineering practice. These margins at the beginning of a mission design phase are typically as follows: new developments: 10–20 %, modified equipment: 5–10 %, unmodified equipment: 5 %.

However, the low cost mission designer should take care in applying margins, as the impact on the mission cost can be negative in the following two scenarios

- Overly conservative margins adopted early in the project: This can lead to an over-design of the system for the mission objectives, or over-paying for a launch if the predicted mass of the spacecraft was high at the time of launch negotiations (e.g. retaining large mass margins).
- Maintaining a standard margin when there is scope to deviate from it with minimal cost implication. For example, a launch is procured with a higher upper limit for mass than is predicted by the spacecraft’s needs. In such a case, it may be possible to make savings in cost/schedule by using this extra mass budget to ease

Table 25.3 Examples of some risks which may be encountered on a project

Example risk	Likelihood	Impact	Mitigation	Level
An EEE component in a unit is obsolete and needs to be replaced with one without flight heritage. May make unit more susceptible to failure	High	Low	Select an alternative component with same functionality. Additional testing can be performed if required (e.g. radiation)	WPM
Externally provided payload equipment on the critical path may suffer delayed delivery. This would cause schedule delay and possible postponement of launch	Low	Medium	Add contingency in project schedule. Look for ways AIT activities can progress before payload arrives (maximise 'platform' level testing)	Project
Resource conflicts within the manufacturing company may lead to another of its missions competing for engineers time, test equipment etc. This could cause major delays to equipment on the critical path of the mission	Medium	High	Prioritisation between the two projects has to be addressed at a corporate level. Mitigation will include recruiting new engineers, or procurement or hire of alternative test equipment and facilities	Corporate

Fig. 25.10 Simplified example of risk management during part of a project

constraints elsewhere in the design; e.g. making module boxes thicker in order to reduce the total radiation dose experienced by the electronic components. This could prevent the need to perform costly radiation testing of new components and hence provide cost savings to the mission without adding any additional risk.

25.4.8.2 Flexibility

The final point to consider when designing a low cost mission relates strongly to the amount of flexibility built into the design. There are many subtle ways that flexibility can provide small cost savings and reduce risk, but there is one major element that can have a large influence on cost: This is to design the spacecraft for launch on a range of launch vehicles. The typical cost for the majority of commercially available launch vehicles is approximately \$40,000 per kg for a typical microsatellite (with a mass less than a few hundred kg) purchasing a dedicated launch will result in an excessively costly mission. The only way to reduce this cost significantly is to share a launch with one or more other satellites. However, finding a launch that fits into the mission schedule and requirements (e.g. the correct orbit) is very challenging. Therefore, in order to ease this problem, the spacecraft should be designed to be compatible with several launch options. This will mean for example that it is qualified to survive the launch environment of all desired options. This is especially important considering that a suitable launch opportunity may not be found until

well into the mission design or manufacture, by which time it will be too late to requalify the structure to a new set of mechanical loads.

25.4.9 Reliability Analysis

In the space industry, it is very important that systems are reliable because (with a few notable exceptions) there is no opportunity to fix hardware once the spacecraft has been launched. Therefore, analysis must be undertaken on the ground to determine exactly how reliable the system is perceived to be. In this context, the term 'reliability' is essentially defined as the probability that the system will satisfy its mission objectives for the required lifetime in its operational environment. Of course, any customer may have a requirement that their system operates for the desired mission lifetime with an acceptable level of availability (freedom from outages) and some may expect to see reliability calculations to 'prove' the system will work in orbit. However, focusing too much effort on up-front reliability predictions can be counter-productive and can increase schedules (and hence cost). The reason for this is that it is very difficult to produce accurate (and hence meaningful) reliability and availability calculations for space missions. Even industry standards to calculate reliability such as MIL-HDBK-217 states the following: "*Predicted and achieved reliability have always been closer for ground electronic systems than avionics*

systems, because the environmental stresses vary less from system to system on the ground.” To take this a step further [3] lists examples where spacecraft have reached the point in the mission where theoretical reliability calculations would predict failure, and yet were still operating nominally, and in some cases for many additional years. Therefore, a low cost approach is to put less emphasis on theoretical predictions and more on experience gained from in-orbit heritage. Reliability analysis can be useful but should not be used as a driver to reduce mission risk or to determine whether a mission will achieve its design lifetime.

So it is important to understand how you actually determine the system is reliable—focusing on a theoretical figure which may not be fully applicable to all aspects can just lead to higher cost with very little benefit to the project. The design of a suitably reliable system need not rely on explicitly calculated figures in order to ensure that a mission will still achieve all of its mission objectives. It is just as valid in many cases to show that in-orbit heritage of the spacecraft and therefore the processes followed throughout the mission design, manufacture, and operation have demonstrated a working system, operating in the actual environment that any similar future mission will experience.

In summary, the risk management for low cost missions should go through the following steps

- Identify all risks.
- Rank them according to likelihood and the severity of their impact on the mission.
- Identify possible fixes and mitigation strategies.
- Focus efforts on reducing those with the biggest potential impact on the ability of the mission to satisfy its mission objectives for the cost and schedule expected. This could be achieved through focused early testing on engineering models, detailed analysis etc.
- For all other risks ask the question: Can the mission live with the impact IF the risk does occur!
- If yes, or it could be fixed at a later date (e.g. in orbit via software upload) then do not spend effort on it now (unless it is part of normal day-to-day work).
- It is not necessary to retire all risks to zero before launch.
- Some things could be an inconvenience or need a work around if they occur, but the mission objectives could still be achieved.

25.4.10 Organizational Structure

As mentioned earlier in the chapter, to achieve a truly low cost mission all aspects of the mission should be addressed, and not just technical but programmatic and even organizational elements. This section focuses on the organizational elements that can help to achieve a truly low cost mission design. This area is highly dependent on the overall

structure of the manufacturing company, and hence it may be difficult for established companies to adopt should they even wish to do so.

Firstly, try to maintain a consistent project team for the full mission cycle. This means that the same engineers are involved in the design, manufacturing, testing, and operational (LEOP at least) phases of the mission. By having separate teams for the design, test, and operations, there will need to be a vast amount of documentation generated at each handover from one team to the other. When the team remains constant this ensures continuity of knowledge throughout the project without the need for vast amounts of time consuming (and hence costly) documentation.

Another programmatic area that is of great importance when designing a low cost mission is to minimize the number of contractual interfaces as far as possible. At a programmatic level, each subcontract has to be managed. This will include attending external meetings (often overseas) and the generation of sufficient documentation by both the prime contractor (to define schedules, requirements etc.) and from the subcontractor (to provide design information and operational instructions, for example). Furthermore, the subcontractor also has to make a profit on their sale, and has to maintain a margin to ensure that they can meet the specification set out for them. All these things lead to a higher cost than if a similar piece of equipment could be produced internally by the prime contractor. An added advantage is that the prime contractor maintains full control over schedule and has complete transparency over the design of internally manufactured equipment. In addition, they also have full control over any risks, giving them the flexibility to implement a pragmatic approach in order to allow focus only on those risks that are deemed most critical to the mission.

Obviously, it will never be possible for a particular satellite manufacturer to make every bit of equipment internally because specialist suppliers are needed for various components and equipment (for example battery supply), but a company that buys the majority of the equipment from external suppliers will find it more difficult to produce a low cost mission.

25.4.11 Mission Operations Impact

So far, the design of a low cost mission has focused primarily on the design of the space segment as well as programmatic elements. However, another way that mission cost can be reduced involves how the spacecraft will be operated. What needs to be identified is some optimal low cost point that is defined by the level of autonomy/automation on the spacecraft and ground segment.

At one extreme, a mission that relays on limited autonomy and hence a large full-time operations team will incur high labor costs. At the other extreme, the entire ground segment and spacecraft is able to operate autonomously. This may appear at first to be the lower cost option, but it is true only if considering the cost of the operational phase of the mission. To design a mission with such a high level of autonomy will require a very costly test campaign on the ground. A vast array of possible failure scenarios must be considered and tested on the ground to ensure the system can recover from these scenarios. This will drive up the costs and schedules of the mission vastly. Therefore, what is desired is a compromise between the two extremes; automate as many of the simple day-to-day operations as possible, while still having operators on call to fix any anomalies that may occur in-orbit.

25.5 Robust Design of Space Systems

In every design phase of an engineering system, component, or process, designers and decision makers need to consider the impact of uncertainty on the budgets and performance of their design solution. A poor consideration for uncertainty would lead to incorrect decisions and to an increased cost in later phases. Starting from the seminal work of Taguchi in the 1950s [1], many scholars have developed computational techniques that aim to achieve an efficient and correct quantification of uncertainty. These techniques can generally be classified depending on the nature of the uncertainty and on the particular context. Uncertainty can be in the design, in the manufacturing, and in the operation of system, process, or component. Uncertainty quantification during operations generally aims at identifying failures based on current knowledge and on probabilistic models of the expected behavior of the system or process. Generally a bottom-up approach is used in which the risk associated with individual component is connected to the risk of all the other components and thence to the status of the system. Uncertainty quantification in manufacturing generally refers to tolerance in the manufacturing process and to the inherent uncertainty of the actual size, mass, and shape of components that affect their performance. Uncertainty in the design process is instead related to the current knowledge of the designers and their subjective judgment. Note that during the design phase, decision makers need to account for manufacturing and operations when they design their system.

Uncertainty exists in two basic forms: aleatoric or epistemic. Aleatoric uncertainty is irreducible and is due to the intrinsic stochastic nature of physical phenomena. It can be well modeled and described with probability distributions and a frequency approach in which the likelihood of an

event depends on the number of occurrences. Probability theory adequately covers aleatoric uncertainty and provides the required mathematical tools to deal with it. Epistemic uncertainty is instead related to the lack of knowledge, and is therefore reducible. It cannot be well modeled with probability distributions (although there are probability-based approaches). Imprecise probability theories are providing the required framework and tools to deal with epistemic uncertainty.

From the distinction made above between uncertainty in design, manufacturing and operations, it can be said that in the early phase of the design of a space mission uncertainty is mainly epistemic, during manufacturing is aleatoric, and during operations a bit of both; though methods based on probability theory are generally used in risk management.

In this section, the focus is on the management of epistemic uncertainty during the early design phase, and its inclusion in the optimization of a space engineering system. The principles of robust design and design for reliability will be introduced together with the associated probability-based techniques. The section will then focus on the use of imprecise probabilities to capture epistemic uncertainty in a robust design and design for reliability. The main interest is to provide a quantification of the design margins and optimize the design solutions under epistemic uncertainty. In other words, the interest is in producing optimal design solutions that are robust against epistemic uncertainties on the input parameters and in producing a correct quantification of the design margins on the system budget and performance, based on current knowledge. The approaches in this section can represent an alternative to the use of ECSS or equivalent standards.

25.5.1 Robust Optimization and Uncertainty Quantification

Concepts of robustness and robust design optimization have been developed independently in different scientific disciplines, mainly in the fields of operations research (OR) and engineering design. The introduction of the concept of robustness in design and manufacturing is generally attributed to Genichi Taguchi, who first proposed a highly influential design philosophy based on the identification and quantification of those noise factors that affect the performance of a product or process [13]. However, the use of Taguchi's approach into an optimization process becomes intractable for even medium dimensional problems. The progressive increase in computing power has stimulated the development of uncertainty quantification and robust design optimization methods in all fields of engineering. Since Taguchi's approach was proposed, many authors have

devised a wide range of methods that are suitable for specific problems. If epistemic uncertainties are not included, the uncertainty in the design parameters can be defined with probabilistic functions. The effect of the uncertainty is then propagated through the system model to compute the mean and the variance of the performance index and constraints. In simple cases in which the model can be treated analytically, the mean and the variance are computed through first and second order expansions of the performance index and constraints. In other cases, Monte-Carlo simulations are used [14]. In some cases, the mean and the variance are then minimized simultaneously with some multiobjective optimization technique [15–17], although in the case that only the feasibility is of interest, only the variance is minimized. Methods for robust optimization using the expected value robustness measure and taking feasibility constraints into account can be found in [18]. Epistemic uncertainties have been treated with fuzzy logic [19] but also possibility theory [20, 21]. More recently, other approaches based on evidence theory have been proposed [22, 23]. An initial effort to apply the principles of robust design optimization to space mission design can be found in the works of Vasile and Bonetti and Vasile [24–26] where the authors applied evidence theory and evolutionary multiobjective optimization to the design of a reusable vehicle, and an aerocapture spacecraft respectively. Later on, Croisard et al. [27–30] applied evidence theory to the design of a low-thrust mission to Mercury. In the same years, Fuchs and Neumaier [31] focused both on the modeling of the uncertainty and on the computational technique to generate robust design solution for multidisciplinary space systems. They used multidimensional potential clouds to model uncertainties and several optimization techniques to solve the optimization under uncertainties.

25.5.2 Robust Optimization and Design for Reliability

The robust design of an engineering system can be formulated in different ways depending on the object of interest. In robust optimization the interest is generally to minimize (maximize) the expected value of one or more design budget f_i , with $i = 1, \dots, m$, and the associated variance (or higher order statistical moments), i.e.

$$\begin{aligned} \min_{\mathbf{d} \in D \wedge \mathbf{u} \in U} \boldsymbol{\mu} &= \{E[f_1(\mathbf{d}, \mathbf{u})], \dots, E[f_m(\mathbf{d}, \mathbf{u})]\} \\ \min_{\mathbf{d} \in D \wedge \mathbf{u} \in U} \boldsymbol{\sigma} &= \left\{E\left[(f_1(\mathbf{d}, \mathbf{u}) - \mu_1)^2\right], \dots, E\left[(f_m(\mathbf{d}, \mathbf{u}) - \mu_m)^2\right]\right\} \end{aligned} \quad (25.6)$$

where \mathbf{d} is the vector of the design or decision variables and \mathbf{u} the vector of the uncertain parameters. This problem is tackled as a weighted sum

$$\min_{\mathbf{d} \in D \wedge \mathbf{u} \in U} \mathbf{w}_\mu^T \boldsymbol{\mu} + \mathbf{w}_\sigma^T \boldsymbol{\sigma} \quad (25.7)$$

or as a constrained optimization problem

$$\min_{\mathbf{d} \in D \wedge \mathbf{u} \in U} \boldsymbol{\mu} \quad \boldsymbol{\sigma} \leq \mathbf{0} \quad (25.8)$$

A conservative approach is to solve the *minmax* problem

$$\min_{\mathbf{d} \in D} \max_{\mathbf{u} \in U} f(\mathbf{d}, \mathbf{u}). \quad (25.9)$$

Although this may appear to be an excessively conservative choice, it has some advantages, as will be illustrated later in this chapter. Note that the design space D and the uncertain space U can overlap in the general case, i.e. $D \cap U \neq \emptyset$, therefore one might want to solve the following modified problem

$$\begin{aligned} \min_{\hat{\mathbf{d}} \in D \cap U} \boldsymbol{\mu} &= \{E[f_1(\hat{\mathbf{d}})], \dots, E[f_m(\hat{\mathbf{d}})]\} \\ \min_{\hat{\mathbf{d}} \in D \cap U} \boldsymbol{\sigma} &= \left\{E\left[(f_1(\hat{\mathbf{d}}) - \mu_1)^2\right], \dots, E\left[(f_m(\hat{\mathbf{d}}) - \mu_m)^2\right]\right\} \end{aligned} \quad (25.10)$$

where now the decision variable has an uncertain component $\hat{\mathbf{d}} = \mathbf{d} + \mathbf{u}$. Robust design optimization is generally distinguished from design for reliability that can be formulated as

$$\begin{aligned} \min_{\mathbf{d} \in D \wedge \mathbf{u} \in U} \{f_1(\mathbf{d}, \mathbf{u}), \dots, f_m(\mathbf{d}, \mathbf{u})\} \\ E(c_1(\mathbf{d}, \mathbf{u})) - \delta_1 \leq 0 \\ \vdots \\ E(c_q(\mathbf{d}, \mathbf{u})) - \delta_q \leq 0 \end{aligned} \quad (25.11)$$

where c_i , with $i = 1, \dots, q$, are some constraints and δ_i some reliability thresholds. Even in this case, the equivalent problem can be considered in which the decision variable has an uncertain component. A recent proposal replaces the computation of the statistical moments with the computation of the cumulative distribution function (CDF) and minimizes the area between the computed CDF and a reference template. Considering the minimum possible value of the design budgets δ_f over $D \times U$, the problem becomes the minimization of the robustness index

$$RI = \int_U |F(\mathbf{d}, \mathbf{u}) - \delta_f| d\mathbf{u}. \quad (25.12)$$

25.5.3 Probability-Based Approaches to Uncertainty Quantification

The general problem that probability-based approaches face is the computation of integrals of the following kind

$$\mu^i(\mathbf{d}) = \int_U f(\mathbf{d}, \mathbf{u})^i P_u(\mathbf{u}) d\mathbf{u} \quad (25.13)$$

which corresponds to the i -th statistical moment, where $f(\mathbf{d}, \mathbf{u})$ is the result of the propagation of the uncertainty through the model or process. Note that \mathbf{d} contains both the spatial and the temporal dimensions, i.e. if a process is time dependent \mathbf{d} contains both the time dependent and time independent components. The random parameter $\mathbf{u}(\omega)$ belongs to the uncertain parameter space U , and $\omega \in \Omega$ is the realization of \mathbf{u} in the probability space (Ω, F, P) where $F \subset 2^\Omega$ is the σ -algebra of events and P is a probability measure. The uncertain parameter \mathbf{u} can have a generic probability density function P_u . Equation 25.13 cannot be computed in closed form in the general case. Approximated expressions using Taylor expansions are possible in some cases. If the integral is computed numerically, a suitable number of values of \mathbf{u} must be propagated through the model and then an appropriate numerical quadrature formula is required to estimate the integral value. The straightforward approach to the problem would use a direct sampling of \mathbf{u} according to the distribution P_u (direct Monte Carlo simulation) and then the integral is computed with, for example, Newton-Cotes quadrature formulas or again through a Monte Carlo approach. The direct sampling of \mathbf{u} and the propagation through the model can be very expensive operations and a large number of samples might be required, which makes the whole process prohibitive.

More advanced techniques to reduce the computational cost have been developed in recent times. The general idea is to replace the exact model response $f(\mathbf{d}, \mathbf{u})$ with a surrogate response $w(\mathbf{d}, \mathbf{u})$ and then sample the surrogate model. In some cases the surrogate model is built using an interpolation function defined on structured grids of integration nodes so that sampling of $w(\mathbf{d}, \mathbf{u})$ is not required. Two popular approaches are mentioned here: generalized polynomial chaos expansions (GPCE) and stochastic collocation (SC). Together they are referred to as *stochastic expansion methods*.

25.5.4 Approaches to Epistemic Uncertainty Quantification

When the applied mathematics literature related to information theory and expert systems is examined, one finds a number of theories that can handle both aleatory and epistemic uncertainty. Some examples are fuzzy set theory, [32–35], interval analysis [36, 37], evidence theory [38–42], possibility theory [43, 44], and theory of upper and lower previsions [45]. Some of these theories only deal with epistemic uncertainty; most deal with both epistemic and aleatory uncertainty; and some deal with other varieties of uncertainty, e.g., logic appropriate for artificial intelligence and data fusion systems. An article by Klir and Smith [46] summarizes how these theories of uncertainty are related to one another from a hierarchical point of view. They show that evidence theory is a generalization of classical probability theory. From the way that evidence theory measures and combines the pieces of evidence supporting one theory (or opinion), it can be considered a generalization of possibility theory, although in evidence theory and in possibility theory the mechanics of operations applied to bodies of evidence are completely different. Note that in some cases evidence theory is referred to as the theory of random sets. Also, the use of imprecise probability theories, interval analysis, or fuzzy set theories is not the only approach to epistemic uncertainty. Many authors proposed the use of probability theory with the assumption of uniform distributions within intervals or p-boxes, defined by intervals, and a probability density function defined on the interval. In the remainder, the focus will be on a possible future use of evidence theory.

25.5.5 Future Perspective: Evidence-Based Robust Design Optimization

Shafer's evidence theory (ET) is a branch of the mathematics of uncertain reasoning that allows the decision-maker to deal with uncertain events and incomplete and conflicting information [47–50]. In ET there are two complementary measures of uncertainty: belief and plausibility, or the lower and upper probabilities that an event can occur. Given all the available pieces of evidence, a single probability distribution cannot be specified; rather, a range of possible probabilities exists, all of which are consistent with the evidence. Belief and plausibility measures can be based on many types of information, e.g. experimental data, theoretical evidence, individual expert opinion, or consensus among experts concerning the range of possible values of a parameter or the possibility of the occurrence of an event. There are two main differences between ET and classical

probability theory: in ET, no probability distribution function needs to be defined and no specific probability needs to be assigned to any value that a variable can assume, the evidence of an event and the evidence of its negation do not sum up to unity. This means that in ET the absence of evidence in support to an event does not imply its negation but leaves the door open to other possibilities.

Evidence theory has been used mainly in information fusion, decision making, and risk analysis. Other applications are in autonomy and intelligent systems, and planning and scheduling under uncertainties. Recently ET has been considered for applications in the robust design of structures and mechanisms in aerospace and civil engineering [22, 23]. It has also received a growing attention for robust design optimization. The main reason is that it can incorporate consistently both epistemic and aleatoric uncertainty in a solid mathematical framework. Furthermore, it is a generalization of both fuzzy sets and possibility theory. Relevant works are connected to the names of Oberkampf and Helton, in particular their seminal work on the possible applications of evidence theory to engineering problems [22], and to the names of Agarwal and Renaud who proposed the use of evidence theory together with response surfaces and gradient methods [23]. Although not specifically intended for engineering applications, relevant works that attempt to make the use of evidence theory efficient have been performed by Tessem [51] and Bauer [52], who proposed different techniques to reduce the set of intervals that must be evaluated in order to compute the belief and plausibility curves. Several other examples of computational techniques exist in the literature that use evidence theory for robust optimization of space systems [24–30, 53]. It was demonstrated that the use of evidence theory leads to more conservative designs than probability theory, at least if the belief function is used as index of the reliability of a design point [24]. Evidence theory has been applied to small to medium problems in structural design and optimization [23, 52]. Recently it was applied to the multiobjective design optimization reusable launchers [24], aerocapture maneuvers [25, 26], and low-thrust trajectories [28–30].

In most current engineering design applications of evidence theory, experts are expected to express their belief on an uncertain parameter u being within a certain set of intervals. Each interval can be considered as an elementary proposition, and all the intervals form the so-called frame of discernment Θ , which is a set of mutually exclusive elementary propositions. The frame of discernment can be viewed as the counterpart of the finite sample space in probability theory. The power set of Θ is $U = 2^\Theta$ or the set of all the subsets of Θ (the uncertain space in the following). The level of confidence an expert has in an element θ of U is quantified using the basic probability assignment (bpa) $m(\theta)$ that satisfies the axioms

$$\begin{aligned} m(\theta) &\geq 0, \forall \theta \in U = 2^\Theta; \\ m(\theta) &= 0, \forall \theta \notin U = 2^\Theta; \\ m(\emptyset) &= 0; \\ \sum_{\theta \in U} m(\theta) &= 1 \end{aligned} \quad (25.14)$$

An element of U that has a non-zero bpa is named a focal element. When more than one parameter is uncertain, the focal elements are the result of the Cartesian product of all the elements of each power set associated with each uncertain parameter. The bpa of a given focal element is then the product of the bpa of all the elements in the power set associated with each parameter. As an example, given the elementary propositions E_1 and E_2 the power set is $U = \{0, E_1, E_2, E_1 \cup E_2\}$ where the disjunctive relation $E_1 \cup E_2$ means that u can be either in E_1 or in E_2 . This last point is quite important as evidence theory quantifies, through $E_1 \cup E_2$, the degree of ignorance. The bpa assignment then becomes

$$m(E_1) + m(E_2) + m(E_1 \cup E_2) = 1. \quad (25.15)$$

The belief Bel and the plausibility Pl functions are defined as follows

$$Bel(A) = \sum_{\forall \theta_i \subseteq A} m(\theta_i); \quad Pl(A) = \sum_{\forall \theta_i \cap A \neq \emptyset} m(\theta_i) \quad (25.16)$$

where A is the proposition about which the belief and plausibility need to be evaluated. For example, the proposition can be expressed as

$$A = \{\mathbf{u} \in U | f(\mathbf{u}) \leq v\} \quad (25.17)$$

where f is the outcome of the system model and the threshold v is the desired value of a design budget (e.g. the mass). Thus, focal elements intercepting the set A but not included in A are considered in Pl but not in Bel .

25.5.5.1 Robust Design Formulations

As in Sect. 25.5.2, a design and optimization problem can be formulated assuming epistemic uncertainties quantified with evidence theory. Consider a function $f : D \times U \rightarrow \mathfrak{R}$ characterizing a system to be optimized, where D is the available design space and U the uncertain space. The function f represents the model of the system budgets (e.g. power budget, mass budget, etc.), and depends on some uncertain parameters \mathbf{u} and design parameters \mathbf{d} such that

$$\mathbf{u} \in U \subseteq \mathfrak{R}^m; \quad \mathbf{d} \in D \subseteq \mathfrak{R}^n. \quad (25.18)$$

A bpa structure is associated with the frame of discernment U of the uncertain parameters \mathbf{u} . From the definition of Bel it is clear that the maximum of f over every focal

element of U should be computed and compared to v . If the maximum and minimum do not occur at one of the vertices of the focal element then an optimization problem has to be solved for every focal element and for each new design vector. Because the number of focal elements increases exponentially with the number of uncertain parameters and associated intervals, so too does the number of optimization problems.

When uncertainty or partial information exists on some inputs to a design process or model (represented by the function f in this section), the interest is to quantify the impact of uncertainty on the outputs (or quantity of interest). If the function f represents a design budget, say the mass of a system, the interest is to quantify the mass margin. Furthermore, the design budget and the margin need to be minimal. If one assumes that a $Bel = 1$ signifies complete confidence under current information and that $Pl = 0$ signifies impossibility under current information, then the total margin can be defined as

$$\Delta_f = v|_{Bel=1} - v|_{Pl=0}. \quad (25.19)$$

Often designers are interested in the variation of the belief with the threshold v , or in other words with the added margin. Indeed, it may be relevant to take a little more risk (a lower value of the belief) if the performance gain is significant. Therefore, it would be interesting to have a complete trade-off curve, solution of the bi-objective optimization problem

$$\begin{aligned} \max_{\mathbf{d} \in D \wedge \mathbf{u} \in U} & Bel(f(\mathbf{d}, \mathbf{u}) < v) \\ \min v & \end{aligned} \quad (25.20)$$

Examples of the solution of this problem can be found in [25–28, 30]. The optimal maximum design margin in Eq. 25.19 can be found by solving the following two optimization problems

$$v|_{Bel=1} = \min_D \max_{\bar{U}} f(\mathbf{d}, \mathbf{u}) \quad (25.21)$$

$$v|_{Pl=0} = \min_D \min_{\bar{U}} f(\mathbf{d}, \mathbf{u}) \quad (25.22)$$

where \bar{U} is the normalized collection of all the focal elements in U . In other words, all the focal elements in U are normalized with respect to the maximum range of the uncertain parameters and collected into a compact unit hypercube in which all the focal elements are adjacent and not overlapping. If one wants to consider disjunctive elements of U , care must be put in the construction of the belief and plausibility values. The mass $m(E_1 \cup E_2)$ must be added to the computation of the belief (respectively the plausibility) if either of the propositions are included in

A (respectively intersects A) but not added twice if both propositions are included in A (respectively intersects A). The unit hypercube is built excluding all disjunctive propositions in U , and then the *bpa* of every partition of \bar{U} is computed, checking whether the partition contains at least one part of a disjunctive proposition. If that is the case, the associated mass is added to the total *bpa* of the partition. If multiple partitions containing the same elements of a disjunctive proposition are added up, only one mass is considered.

Other authors focused more on the design for reliability assuming epistemic uncertainty. In this case, the problem can be formulated as follows. Let us assume that one has to solve the following constraint minimization problem

$$\begin{aligned} \min_{\mathbf{d} \in D} & f(\mathbf{d}) \\ \mathbf{c}(\mathbf{d}, \mathbf{u}) & \leq 0. \end{aligned} \quad (25.23)$$

If the interest is to maximize the evidence that the constraints are satisfied also under uncertainty then the problem can be formulated as follows

$$\begin{aligned} \min_{\mathbf{d} \in D} & f(\mathbf{d}) \\ \min_{\mathbf{d} \in D} & Pl(\mathbf{c}(\mathbf{d}, \mathbf{u}) > \varepsilon) \end{aligned} \quad (25.24)$$

where ε is an acceptable constraint violation. For an efficient solution of this problem see Zhou et al. [54]. One can also combine robust optimization and reliability maximization in the following integrated formulation [26]

$$\begin{aligned} \max_{\mathbf{d} \in D} & Bel(f(\mathbf{d}, \mathbf{u}) < v) \\ \min v & \\ \min_{\mathbf{d} \in D} & Pl(\mathbf{c}(\mathbf{d}, \mathbf{u}) > \varepsilon) \end{aligned} \quad (25.25)$$

25.5.6 Application Example: Evidence Based Telecommunication System Design

This section describes an example of designing a telecom system assuming that there are uncertainties on some key parameters. The example is used to illustrate the different results that can be obtained by applying the ECSS recommendation and a more rigorous ET-based quantification of the margins.

Assume that the mass of the telecommunication system depends on the link budget and on the mass of the electronics only. Furthermore, assume that the link budget depends only on the following parameters: η_{ANT} is the efficiency of the antenna, L_r is the line losses, T_{ant} is the temperature of the antenna, f_T is the central frequency of the carrier, Mod is the type of modulation, T is the type of amplifier, and G_T is

Table 25.4 TTC bba structure

η_{ANT}	Interval	[0.5 0.6]	[0.65 0.75]	[0.6 0.8]	[0.8 0.95]
	bba	0.2	0.5	0.2	0.1
ρ_{CMR}	Interval	[0.1 0.2]	[0.25 0.3]	[0.1 0.3]	
	bba	0.5	0.35	0.15	
L_t	Interval	[1 2]	[2 3]	[3 5]	
	bba	0.2	0.3	0.5	
T_{ant}	Interval	[200 250]	[300 370]	[400 500]	
	bba	0.1	0.6	0.3	

Table 25.5 Design space for TTC

Parameter	Low bound	Upper bound
f_T (MHz)	7e3	11e3
Mod	0	1
T	0	1
G_T (dB)	5	20

Fig. 25.11 Bel and Pl curves for the TTC system: comparison between margin estimation and evidence in the optimistic case

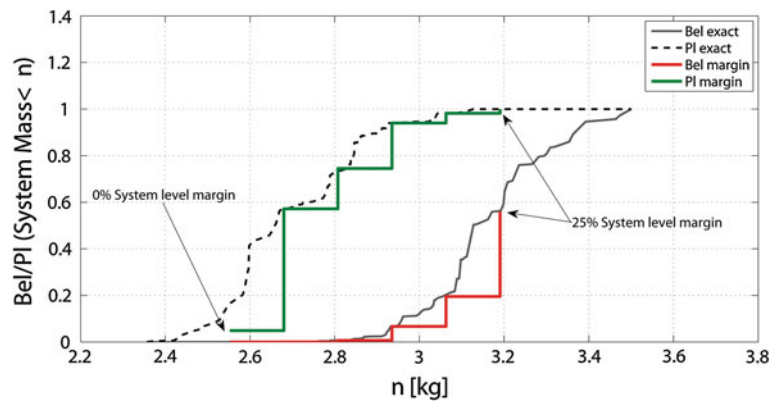
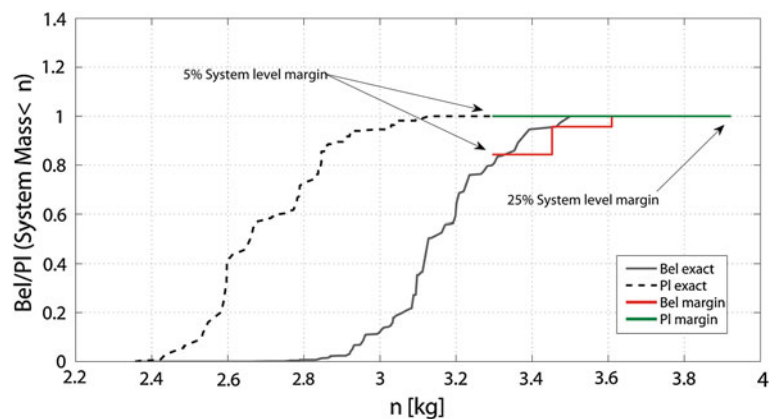


Fig. 25.12 Bel and Pl curves for the TTC system: comparison between margin estimation and evidence in the conservative case



the gain of the transmitting antenna. The mass of the electronics depends on the density parameter ρ_{CMR} .

Now assume that η_{ANT} , L_t , T_{ant} and ρ_{CMR} are uncertain at the beginning of the design process (say in a typical concurrent design session), while f_T , Mod , T and G_T can be

controlled in order to minimize the mass of the telecommunication system and the associated margin due to uncertainty, i.e. they are the design parameters. Table 25.4 summarizes the intervals of variability of the uncertain parameter and the associated *bpa*, namely how much

confidence the domain experts have in the values that the uncertain parameters can assume. Table 25.5 shows instead the range of variability of the design parameters.

The *Bel* and *Pl* curves computed with an exact quantification of the uncertainty, and the *Bel* and *Pl* associated with the system mass as computed using a margin approach based on experience and the ECSS standards are shown in Fig. 25.11. It is assumed, in this case, that a designer takes the best possible value for the mass of the system, $v|_{Pl=0}$, and the associated design solution \mathbf{d}_{\min} and adds a 25 % margin to the computed power requirement and also to the mass of the casing of the electronics. Then the designer adds between 0 and 25 %, with 5 % increments, to the overall mass of the system and computes the associated *Bel* and *Pl*. This gives a measure of the credibility of the designer's use of the margins. Figure 25.11 demonstrates that even in the case of a 25 % margin at the system level the margin approach based on experience is not able to capture the actual content of uncertainty, reaching less than 0.6 belief in the mass of the system. Now if one assumes a more conservative approach, in which a 25 % margin is added also to the mass of the antenna and the mass of the amplifier, the designer will get the result in Fig. 25.12. In this case the expectation of the designer is far too conservative, overestimating the maximum system mass.

The *Bel* margin curve in Fig. 25.12 is computed taking the min/min solution and adding 25 % margin to the link power P_{Ld} , to the antenna mass, and to the amplifier mass. An additional margin is added to the whole system mass. The belief is computed for different values of the overall margin ranging from 5 to 25 %. The system mass for the maximum overall margin is worse than the min/max solution. On the other hand, if the overall system margin is reduced below 25 % the belief drops rapidly. Note that the calculation of the actual reliability of the margin solution would not be possible without the use of evidence theory.

References

1. R.H.Battin, An introduction to the mathematics and methods of Astrodynamics, AIAA Education series, AIAA, New York, 1987
2. S.Kemble Interplanetary Mission Analysis and Design: Springer Praxis 2006
3. J.R,Wertz, Orbit & Constellation Design & Management, Microcosm, 2001
4. J. A. Aguilar, A. B. Dawdy, G. W. Law, Aerospace Corporation's Concept Design Center, Proceedings of the 8th Annual International Symposium of the International Council on Systems Engineering, July 26-30, 1998
5. G. E. COOK, SATELLITE DRAG COEFFICIENTS, Planet. Space Sci. 1965, Vol. 13, pp. 929 to 946. Pergamon Press Ltd.
6. ECSS-Q-ST-60-C on Electrical, Electronic and Electromechanical (EEE) components, rev.1, March 2009
7. International Council of System Engineering, System Engineering Terms Glossary, www.incose.org, date: Oct-1998
8. D. K. Sobek, A. C. Ward, Principles from Toyota's set-based concurrent engineering process, the 1996 ASME Design Engineering Technical Conferences and Computers in Engineering Conference
9. <http://jplteamx.jpl.nasa.gov/>
10. M. Bandecchi, B. Melton, B. Gardini, F. Ongaro, The ESA/ ESTEC Concurrent Design Facility, EuSEC 2000
11. R. Cook, G. Kazz, W. Tai, The Mars Pathfinder End-to-end information system – A Pathfinder for the development of future NASA planetary missions, SpaceOps '96, Proceedings of the Fourth International Symposium held 16-20 September 1996 in Munich, Germany
12. CDF System Description, ESA internal document CDF-SYS-001, 20 January 2008
13. Taguchi G., Quality Engineering through Design Optimization, Kraus International Publications, New York, 1984.
14. Du X., Chen W. Towards a better understanding of modeling feasibility robustness in engineering design, ASME J. Mech. De. 122 (2) (2000) 291-311.
15. Chen W., Wiecek M., Zhang J., Quality utility- a compromise programming approach to robust design, ASME J. Mech. De. 121 (2) (1999) 179-187.
16. N. Rolander, J. Rambo, Y. Joshi, J. Allen, F. Mistree, An approach to robust design of turbulent convective systems, J. Mech. Des. 128 (4) (2006) 844-855.
17. Y. Jin, B. Sendhoff, Trade-off between performance and robustness: an evolutionary multiobjective approach, in C. Fonseca, P. Fleming, E. Zitzler, K. Deb (Eds.), Evolutionary Multi-Criterion Optimization: Second International Conference, EMO 2003, Springer-Verlag, Hidelberg, 2003, pp. 237-251.
18. Sundarsen S. Ishii K. Houser D., A robust optimization procedure with variations on design variables and constraints, in : ASME Design Automation Conference, ASME, 1993, pp. 387-394.
19. Arakawa M., Yamakawa H., Ishikawa H., Robust design using fuzzy numbers with intermediate variables, in: 3rd World Congress of Structural and Multidisciplinary Optimization, 1999.
20. Choi L., Amd Du K.K., Youn B., Gorsich D., Possibility-based design optimization method for design problems with both statistical and fuzzy input data, in : 6th World Congress of Structural and Multidisciplinary Optimization, Rio de Janeiro, Brazil, 2005.
21. Youn B., Choi L., Amd Du K.K., Gorsich D., Integration of possibility-based optimization to robust design for epistemic uncertainty, in : 6th World Congress of Structural and Multidisciplinary Optimization, Rio de Janeiro, Brazil, 2005.
22. Oberkampf W.L. Helton J.C. Investigation of Evidence Theory for Engineering Applications. AIAA 2002-1569, 4th Non-Deterministic Approaches Forum, 22-25 April 2002, Denver Colorado.
23. Agarwal H., Renaud J.E., Preston E.L. Trust Region Managed Reliability Based Design Optimization using Evidence Theory. AIAA 2003-1779, 44th AIAA/ASCE/AHA Structures, Structural Dynamics and Materials Conference, 7-10 April 2003, Norfolk, Virginia.
24. Vasile M., Bonetti D. Evolution of the Concurrent Design Process Under Uncertainties. International Concurrent Engineering Workshop, ESA/ESTEC 30 September-1 October 2004.
25. Vasile M. Robustness Optimisation of Aerocapture Trajectories Design Using a Hybrid Co-evolutionary Approach. 18th

- International Symposium on Spaceflight Dynamics. 11-15 October 2004, Munich, Germany.
26. Vasile M. Robust mission design through evidence theory and multiagent collaborative search. *Annals of the New York Academy of Sciences*, 1065:152–173, December 2005.
 27. Croisard, N., Ceriotti, M., Vasile, M., Uncertainty Modelling in Reliable Preliminary Space Mission Design (extended abstract), Workshop on Artificial Intelligence for Space Applications (IJCAI-07), Hyderabad, India, January 2007.
 28. Croisard N., Vasile M., Kemble S., Radice G., Preliminary Space Mission Design Under Uncertainty. IAC-08-D1.3, Glasgow 2008.
 29. Croisard N., Vasile M., Kemble S., Radice G., Preliminary Space Mission Design Under Uncertainty, *Acta Astronautica*, 2009, doi:10.1016/j.actaastro.2009.08.004.
 30. Croisard N., Vasile M., System Engineering Design Optimisation Under Uncertainty for Preliminary Space Mission. IEEE Congress on Evolutionary Computation 2009, 18th-21st May, 2009, Trondheim, Norway.
 31. M. Fuchs and A. Neumaier, Handling uncertainty in higher dimensions with potential clouds towards robust design optimization, pp. 376-382 in: *Soft Methods for Handling Variability and Imprecision* (D. Dubois et al., eds.), *Advances in Soft Computing*, Vol. 48, Springer 2008.
 32. Manton, K. G., Woodbury, M. A., and Tolley, H. D., *Statistical Applications Using Fuzzy Sets*, John Wiley, New York, 1994.
 33. Onisawa, T., and Kacprzyk, J., eds. *Reliability and Safety Analyses Under Fuzziness*, Physica-Verlag Heidelberg, 1995.
 34. Klir, G. J., St. Clair, U., and Yuan, B., *Fuzzy Set Theory: Foundations and Applications*, Prentice Hall PTR, Upper Saddle River, NJ, 1997.
 35. Dubois, D., and Prade, H., eds. *Fundamentals of Fuzzy Sets*, Kluwer Academic Publishers, Boston, MA, 2000.
 36. Moore, R. E., *Methods and Applications of Interval Analysis*, SAIM, Philadelphia, PA, 1979.
 37. Kearfott, R. B., and Kreinovich, V., eds. *Applications of Interval Computations*, Kluwer Academic Pub., Boston, MA, 1996.
 38. Guan, J., and Bell, D. A., *Evidence Theory and Its Applications*, Vol. I, North Holland, Amsterdam, 1991.
 39. Krause, P., and Clark, D., *Representing Uncertain Knowledge: An Artificial Intelligence Approach*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
 40. Kohlas, J., and Monney, P.-A., *A Mathematical Theory of Hints - An Approach to the Dempster- Shafer Theory of Evidence*, Springer, Berlin, 1995.
 41. Klir, G. J., and Wierman, M. J., *Uncertainty-Based Information: Elements of Generalized Information Theory*, Vol. 15, Physica-Verlag, Heidelberg, 1998.
 42. Kramosil, I., *Probabilistic Analysis of Belief Functions*, Kluwer, New York, 2001.
 43. Dubois, D., and Prade, H., *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York, 1988.
 44. De Cooman, G., Ruan, D., and Kerre, E. E., eds. *Foundations and Applications of Possibility Theory*, World Scientific Publishing Co., Singapore, 1995.
 45. Walley, P., *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.
 46. Klir, G. J., and Smith, R. M., On Measuring Uncertainty and Uncertainty-Based Information: Recent Developments, *Annals of Mathematics and Artificial Intelligence*, Vol. 32, No. 1-4, 2001, pp. 5-33.
 47. Dempster A.P. (1967): "Upper and Lower Probabilities Induced by a Multivalued Mapping", *The Annals of Mathematical Statistics*, 38, pp. 325-338.
 48. Shafer G. (1976): *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.
 49. Shafer G. (1990): "Perspectives on the Theory and Practice of Belief Functions", *International Journal of Approximate Reasoning*, 4, pp. 323-362.
 50. Zadeh, L., "Review of Shafer's A Mathematical Theory of Evidence," *Artificial Intelligence Magazine*, Vol. 5, 1984, pp. 81–83.
 51. Tessem B. Approximation for efficient computation in the theory of evidence. *Artificial Intelligence* 61 (1993) 315-329, Elsevier.
 52. Bauer M. Approximation for Decision Making in the Dempster-Shafer Theory of Evidence. In *Uncertainty in Artificial Intelligence*, 1996, 73–80, Morgan Kaufmann Publishers.
 53. Vasile M., Robust Optimization of Trajectory Intercepting Dangerous NEO. AAS/AIAA Astrodynamic Specialist Conference, 5-8 August 2002, Monterey, California, U.S.A.
 54. Zhou, Jun, and Zissimos P. Mourelatos, "A sequential algorithm for possibility-based design optimization," *Journal of Mechanical Design*, Volume 130, January 2008.

Index

A

- Ablator materials, [101](#)
 - shielding, [101](#), [106](#)
- Absorptance
 - spectral directional, [377](#)
 - spectral hemispherical, [377](#)
- Accelerometer, [359](#)
- Active pixel sensors (APS). *See* Sun sensor
- Ada. *See* Programming language
- Adams, John Couch, [75](#)
- ADCS. *See* Attitude determination and control system
- Adiabatic diamagnetic refrigerators (ADR), [140](#)
- Advanced orbiting service (AOS), [481](#)
- Advancement degree of difficulty, [34](#)
- Aeroassist, [96](#)
- Aerobraking, [96](#), [102](#)
- Aerodynamic efficiency
 - A6, [173](#), [174](#)
 - K1, [173](#)
- Aerodynamics, [102](#), [109](#), [167](#), [174](#), [537](#)
 - control surface. *See* Control surface
- Aerology, [174](#)
- Aeronautics, [25](#)
- Aerospace, [25](#)
- Aggregat series (rocket), [3](#)
- Agriculture (in space), [501](#)
- Airbag, [204](#), [533](#), [547](#)
- Albedo, [379](#)
 - bond, [88](#)
 - geometric, [88](#)
- Albert II (monkey), [8](#)
- Alloys (metallic), [220](#), [222](#), [238](#)
- Altimeter, [355](#), [368](#)
- Amplifier
 - Gallium Nitride Solid-state power amplifier (GaN SSPA), [436](#)
 - high-power amplifier (HPA), [414](#)
 - low noise, [413](#)
 - solid-state power amplifier (SSPA), [414](#)
- Angular momentum, [64](#)
- Anik (spacecraft), [398](#)
- Annihilation, [281](#)
- Antenna, [402](#), [413](#), [416](#), [418](#)
 - biconical, [420](#)
 - cassegrain, [421](#), [425](#)
 - coverage polygon, [420](#)
 - cross-pol discrimination (XPD), [419](#)
 - directivity, [419](#)
 - dual gridded reflector (DGR), [421](#)
 - effective aperture area, [420](#)
 - feed array, [423](#)
 - gain, [420](#)
 - gregorian, [421](#), [425](#)
 - isotropic, [410](#)
 - large offset antenna demonstrator, [244](#)
 - lens, [422](#)
 - maximum achievable directivity, [420](#)
 - noise temperature, [402](#)
 - phased array, [423](#)
 - pointing mechanisms (APMs). *See* Mechanisms, pointing
 - polarization, [419](#)
 - reflectarray, [437](#)
 - reflector, [420](#)
 - shaped beam, [423](#)
 - shaped surface reflector, [424](#)
 - sidelobe, [421](#)
 - unfurlable reflector, [422](#)
- AOCS. *See* Attitude control
- Aphelion. *See* Apoapsis
- Apoapsis, [64](#)
- Apocentre. *See* Apoapsis
- Apogee. *See* Apoapsis
- Apollo program, [8](#), [101](#), [103](#), [204](#), [545](#), [619](#), [658](#)
 - command module, [517](#), [520](#)
 - guidance computer (AGC), [471](#)
 - lunar module, [531](#)
 - lunar surface access module, [289](#)
 - moon buggy, [545](#)
- Argument of latitude, [67](#)
- Argument of periapsis, [66](#)
- Ariane 5 (rocket), [168](#), [171](#), [172](#), [174](#), [202](#), [284](#)
- Aristotle, [61](#)
- ARPANet, [21](#)
- Arrhenius law, [113](#)
- Artemis (spacecraft), [228](#)
- ASTER (optical sensor), [16](#)
- Asteroid, [365](#)

A (cont.)

Astrodynamics
 definition of, 61
 Astronaut, 16
 Astronautics, 25
 Astronomical unit (au), 40
 Astronomy, 17
 Atmosphere, 38
 aerodynamic disturbance, 338
 AIAA guide to reference and standard
 atmosphere models, 107
 atmospheric drag. *See* Disturbing forces, atmospheric
 climate, 106, 114
 density profile, 39, 536
 Earth, 83, 106, 537
 entry, descent, and landing, 517
 global circulation, 108
 greenhouse effect, 107
 Ionosphere, 41
 lapse rate, 107
 Mars, 106, 537
 Mercury, 107
 models, 39, 82, 83, 536
 Global reference atmosphere models (GRAM), 83, 537
 US standard atmosphere, 83
 orbit decay time, 86
 temperature profile, 39, 536
 Titan, 537
 Venus, 106, 537
 weather, 106, 108
 Atmospheric entry vehicle, 99, 100, 109, 203, 515, 518
 aerodynamics, 520
 Atmospheric Reentry Demonstrator (ARD), 203, 520
 constraints, 519
 definition of, 515
 European EXPERimental Reentry Testbed (EXPERT), 203
 geometry, 520
 heating, 519
 inflatable reentry and descent
 technology (IRDT), 203, 521
 Intermediate eXperimental Vehicle (IXV), 203
 trajectory, 521
 Atmospheric glow, 40, 49, 50, 126
 Atomic oxygen (ATOX). *See* Oxygen, erosion
 Attitude control, 134, 333, 335
 actuators, 344
 magnetic, 346
 thrusters, 348
 Attitude determination, 328, 335
 algorithm, 328
 recursive, 330
 sensors, 338
 Attitude stabilization
 bias momentum, 336
 dual spin, 336
 gravity gradient, 335
 magnetic, 336
 reentry, 103, 104
 spin, 104, 336
 zero momentum, 337
 Aurora, 40, 42, 51
 Automated Transfer Vehicle (ATV) (spacecraft), 348
 Automatic code generation, 477
 Automatic Threshold Adjust (ATA). *See* Sun sensor
 Autonomy, 127, 130, 352, 449, 480, 481, 541, 707
 Azimuth, 70

B

Babakin, Georgy, 533
 Baikonur cosmodrome, 4, 169
 Ballistic
 coefficient, 84, 100, 102, 518
 trajectory, 102
 Balloon, 532
 Bandwidth, 417
 Bartoli, Adolfo, 86
 Bartz equation. *See* Nusselt relation
 Barycenter. *See* Coordinate systems
 Baseline mission, 131
 Battery, 255
 cycle life, 256
 depth of discharge (DoD), 256
 energy density, 256
 figures of merit, 256
 gravimetric energy density, 256
 life, 257
 performance, 256
 power density, 256
 reconditioning, 595
 safety considerations, 256
 specific energy, 256
 specific power, 256
 state of charge (SoC), 256
 type, 256
 Beamed power, 273
 Beam forming network (BFN), 423
 BeiDou navigation system, 15
 Bent-pipe (communications) payload. *See* Communication system,
 non-processing payload
 BepiColombo (mission), 7, 39, 99, 235
 Bi-elliptic transfer, 93
 BILSAT-1 (spacecraft), 346
 Biological environment, 498
 Bipropellant. *See* Propellant
 Bit error rate (BER), 408, 412
 Black-body
 planetary radiation. *See* Planetary radiation
 radiation, 376
 solar spectrum. *See* Solar spectrum
 Blazing arrow, 187
 Bode's law, 62
 Body-fixed co-ordinate system (ITRF). *See* Coordinate systems
 Bolometer, 341
 Boltzmann
 constant, 376, 403
 equation, 110
 Bond, George Phillips, 76
 Bond, William Cranch, 76
 Bonding. *See* Joining
 Bosch process, 500
 Bose, Raj Chandra, 409
 Boundary element analysis (BEA), 209
 Boundary layer, 105
 Bow shock, 110
 Brahe, Tycho, 61
 Brayton cycle. *See* Power
 Brazing. *See* Joining
 Breadboard, 602
 Bright Star Catalog (BSC), 342
 Broadcasting, 401
 Buran (spacecraft), 101, 520
 Buried charging. *See* Internal electrostatic charging/discharging
 (IESD)

- C**
- Cabin**
- acoustics, 495
 - air, 493
 - dust control, 500
 - heating, 507
 - illumination, 494
 - trace contaminant, 500
- Calendar, 71
- Calorimeter, 139
- Camera. *See* Optical imaging instrument
- CAN. *See* Controller area network (CAN)
- Canadarm, 542, 543
- Capability Maturity Model Integration (CMMI), 477
- Cape canaveral (Air force station), 5, 169
- Čapek, Karel, 541
- Carbon fiber reinforced plastics/polymers (CFRP), 221
- Carnot cycle efficiency. *See* Power
- Carrier-to-interference ratio, 410
- Carrier-to-noise ratio, 410
- Cascaded network, 403
- Cassini-Huygens (mission), 37, 55, 95, 281
 - flight software, 472
 - huygens probe, 100, 520
- Cavitation, 186
- Celestial mechanics, 61
- Celestial sphere, 69
- Centrifuge, 215
- Ceramics, 222
- CeSiC®. *See* Ceramics
- CFD. *See* Computational fluid dynamics
- Chang-Díaz, Franklin, 312
- Change control board (CCB), 645
- Channel access data unit (CADU), 545
- Channel amplifier (CAMP), 432
- Characteristic velocity, 178, 286
- Charge couple device (CCD), 451
- Chaudhuri, Dwijendra Kumar Ray-, 409
- Child-Langmuir law, 305
- Cholesterol, 506
- Chugging. *See* Rocket engine, combustion instabilities
- Circadian biorhythm, 495
- City lights. *See* Stray light
- Clamp band, 198
- Clarke, Arthur C., 397
- Classical orbital elements, 66
- Clementine (spacecraft). *See* Deep Space Program Science Experiment (DSPSE)
- Clohessy-Wiltshire equations of motion, 97, 361, 364
- Clothes, 507
- Coarse pointing assembly (CPA). *See* Mechanisms, pointing
- Collected volatile condensable material (CVCMM), 208
- Colony formation unit (CFU), 494
- Columbus laboratory. *See* International space station
- Comet halley, 548
- Command link control word (CLCW), 444
- Command pulse distribution unit (CPDU), 443
- Common extensible cryogenic engine (CECE), 295
- Communication system, 397, 688
 - antenna. *See* Antenna
 - architecture, 399, 413
 - bandwidth. *See* Bandwidth
 - input filter assembly. *See* Input filter assembly
 - input section, 413
 - link design, 410
 - noise from external sources, 403
 - noise temperature, 402
 - antenna, 402–404
 - atmospheric brightness, 403
 - brightness, 403
 - Earth, 403
 - effective, 402
 - normal ambient, 403
 - sky brightness, 403
 - non-processing payload, 399, 412, 413
 - output section, 414
 - processing payload, 400
 - receive section. *See* Communication system, input section
 - receive system. *See* Receiver
 - receiver assembly, 426
 - transceiver software, 473
 - transmission media, 404
 - transmit section. *See* Communication system, output section
 - transmit system, 402
- Communications operation procedure (COP), 443
- Communications technology satellite ('Hermes'), 398
- Compass (spacecraft). *See* BeiDou navigation system
- Composite, 167, 220, 224
 - drape, 224
 - matrix, 220
 - tack, 224
- Computational fluid dynamics, 109, 175
- Comstar (spacecraft), 398
- Concept of operations, 151
- Concurrent
 - design, 694
 - engineering, 692
- Conductive heat transfer, 374
 - meshing, 374
 - nodes, 374
- Configuration factor. *See* View factors
- Conductive shape factors, 374
- Configuration management, 157, 477
 - redlining, 158
- Conic section, 63
- ConOps. *See* Concept of Operations
- Constellation
 - definition of, 26
- Consultative committee for space data systems (CCSDS), 443
- Contamination, 48
 - control, 126
 - detectors, 121
 - electromagnetic, 121
 - measurement, 138
- Contingency, 148
 - anticipated contingency factor, 148
 - critical resource, 148
 - growth factor, 148
 - philosophy, 705
- Continuous risk management (CRM). *See* Risk, management
- Continuous thrust, 94
- Continuum (flow), 109, 110, 112
- Contour (spacecraft), 548
- Control moment gyroscopes (CMG), 226, 346
- Control surface
 - aerodynamic, 204, 359
 - thermal, 125, 126, 380
- Controller
 - closed-loop, 359
 - finite-horizon optimal control, 360
 - linear time-invariant (LTI), 360
 - linear-quadratic regulator (LQR), 360

- multiple-input multiple-output (MIMO), 360
 - proportional-derivative (PD), 333
 - proportional-integral-derivative (PID), 333, 360
 - quaternion feedback, 333
 - single-input single-output (SISO), 360
 - state-feedback, 334
 - terminal, 360
 - zero-order hold, 360
 - Controller area network (CAN), 457
 - arbitration, 459
 - CANopen, 460
 - data frame, 458
 - error frame, 459
 - Coordinate systems, 67
 - barycenter, 67
 - body-fixed coordinate system (ITRF), 70
 - earth centered inertial, 67
 - entry, descent and landing, 536
 - Gaussian coordinate system, 70
 - geocentric celestial coordinate system (GCRF), 70
 - geocentric equatorial coordinate, 67
 - geocentric inertial, 67
 - heliocentric coordinate system, 67
 - heliocentric inertial, 67
 - International Celestial Reference System (ICRS), 69
 - satellite-based, 70
 - synodic system, 67
 - Coquilhat, Casimir Erasme, 2
 - Corona (spacecraft), 101
 - Coronal mass ejection (CME), 43
 - Cosmic rays. *See* Galactic cosmic ray (GCR)
 - Cosmonaut, 16
 - Cosmos-1 (spacecraft), 274
 - COSMO-SkyMed, 19
 - Cospas-Sarsat, 13, 15
 - Cost, 56, 160
 - account manager (CAM), 639
 - actual cost of work performed (ACWP), 161
 - basis of estimate (BOE), 639
 - budget at completion (BAC), 161
 - budgeted cost of work scheduled (BCWS), 161
 - budgeting, 639
 - cost performance index (CPI), 161
 - design to cost, 160
 - earned value management, 161, 610
 - estimate, 162
 - estimate at completion (EAC), 161, 641
 - estimation, 162, 636, 639
 - variance (CV), 161
 - workforce, 639
 - Covariance matrix, 172
 - Cowell, Philip Herbert, 76
 - Crane model. *See* Simulation, rain attenuation
 - Crank-Nicholson method, 392
 - Crew psychology, 494
 - Critical inclination (orbit), 91
 - Critical resource. *See* Contingency
 - Crommelin, Andrew Clause de la Cherois, 76
 - Cryogenic system, 387
 - cooling, 387
 - payload, 139
 - rocket engine, 289
 - CryoSat-2 (spacecraft), 30
 - CubeSat, 7
 - Cupola. *See* International space station
 - Curiosity rover. *See* Mars Science Laboratory (spacecraft)
- D**
- Damping, 209, 218
 - active, 219
 - modal viscous, 209
 - Data compression, 133, 453, 484
 - Data routing, 452, 465
 - controller area network (CAN). *See* Controller area network (CAN)
 - interconnections, 400
 - MIL-STD-1553. *See* MIL-STD-1553
 - serial links, 483
 - spaceFibre, 468
 - spaceWire. *See* SpaceWire
 - Data system
 - architecture, 441
 - function, 441
 - network, 446
 - payload, 450
 - storage, 448, 466
 - Dawn (spacecraft), 138, 548
 - Debye
 - length, 46
 - shielding, 46
 - Decompression syndrome, 493
 - Deep impact (spacecraft), 365, 548
 - Deep Space Network (DSN), 29, 597
 - Deep Space Program Science Experiment (DSPSE), 37, 45
 - Deep Space-1 (DS1) (spacecraft), 138, 352, 548
 - Deep Space-2 (spacecraft), 520
 - Defense Meteorological Satellite Program (DMSP), 19, 51
 - Defense support program satellites, 20
 - Deployable structures, 238
 - accuracy/stability, 239
 - actuation, 239
 - articulated foldable masts, 241
 - bi-dimensional deployment appendages, 241
 - categories of, 240
 - coilable masts, 241
 - collapsible tube mast (CTM), 241
 - Harris Hoop-truss, 242
 - inflatable, 242
 - large antenna, 242
 - large reflectors, 242
 - reliability, 239
 - single deployment appendages, 240
 - stiffness, 239
 - storable tubular extendable member (STEM), 241
 - telescopic masts, 241
 - uni-dimensional deployment appendages, 241
 - verification. *See* Verification & validation
 - Deployment mechanisms. *See* Mechanisms
 - Depressurization, 209
 - Descent, 515, 522
 - definition of, 515
 - parachute. *See* Parachute
 - retro-rockets, 532
 - transverse impulse rocket system (TIRS), 532
 - Design
 - 80/20 rule, 697
 - evolutionary development, 700
 - robust, 708
 - spacecraft, 687
 - system, 687
 - Development constraints, 166
 - Dextre, 543
 - Dichroic surface, 421
 - Diffuse surface, 375

- Digital beam forming (DBF), 439
 - Digital video broadcasting by satellites (DVB-S), 400, 409
 - Direction cosine matrix, 323
 - Direct-simulation Monte Carlo (DSMC). *See* Simulation
 - Direct-to-home (DTH) broadcast, 5, 398
 - Discharge coefficient, 178
 - Dissociation, 105, 109, 113, 205, 308, 519
 - Disturbing forces, 80, 337
 - atmospheric, 82, 85, 338, 537, 689
 - drag area, 523
 - general relativity (GR), 62, 88
 - gravitational perturbations, 80, 337, 537, 689
 - internal torques, 338
 - non-spherical central body, 81
 - radiation pressure, 86
 - solar (SRP), 87, 317, 337
 - solar wind, 88
 - tides, 88
 - DODGE (spacecraft). *See* US Department of Defense Gravity Experiment (DODGE) spacecraft
 - Doppler radar, 355
 - Dose (radiation), 45
 - Down converter (communications), 413, 426
 - Drag coefficient, 85, 102, 167, 174, 523
 - Dragon capsule (spacecraft), 520
 - DSMC. *See* Simulation, direct-simulation Monte Carlo
 - Dust, 44, 126
 - dusty plasmas, 44
 - Duct overpressure (DOP), 175
 - Dutch roll, 104
 - Dynamic
 - envelope, 209
 - response, 209
 - Dynamic pressure, 167
 - buffeting, 174
 - maximum, 170
 - Dynamic test loads. *See* Verification & validation, dynamic test loads
 - Dyno (project). *See* GRAB-1 (spacecraft)
- E**
- Early bird (spacecraft). *See* Intelsat-1
 - Earth observation, 15
 - Earth science, 17
 - Earth sensor, 340
 - Eccentric anomaly, 66
 - Eccentricity, 64
 - ECSS. *See* European cooperation for space standardization
 - ERDS. *See* European Data Relay Satellite (ERDS)
 - EDUSAT (spacecraft), 13
 - Efficiency coefficient, 178
 - Einstein, Albert, 62
 - Eisenhower, Dwight D., 8
 - Ekran (spacecraft), 5
 - Electric fields, 39
 - Electric propulsion. *See* Propulsion
 - Electric propulsion pointing mechanism (EPPM). *See* Mechanisms, pointing
 - Electrical motor, 226
 - magnetostrictive, 227
 - Piezo, 227
 - Electro mechanical actuator (EMA), 233
 - Electromagnetic
 - compatibility (EMC), 50, 52, 55, 265
 - interference (EMI), 52, 55, 265
 - interference sources, 265
 - interference suppression, 265
 - radiation, 86
 - Electronic parts
 - class B, 52
 - class S, 52
 - commercial off-the-shelf (COTS), 45, 702
 - de-rating for reliability, 266
 - nuclear hardened, 45
 - preferred parts list, 266
 - rad-hard, 45
 - space-qualified, 45
 - Electrostatic charging/discharging
 - electrostatic discharge (ESD), 52
 - Element set, 66, 166
 - Elevation, 70
 - Emissivity
 - spectral directional, 377
 - spectral hemispherical, 377
 - Encke, Johann Franz, 76
 - Engineering Test Satellite-VII (ETS-VII), 544
 - Entry. *See* Atmospheric entry vehicle
 - corridor, 102
 - velocity, 518
 - Entry, descent and landing systems (EDLS), 516
 - coordinate system. *See* Coordinate systems
 - design, 517
 - simulation. *See* Simulation
 - ENVISAT—Environmental satellite (spacecraft), 26
 - Epoch, 71
 - Equinoctial elements, 78
 - modified, 78
 - Error, 449
 - detection and correction (EDAC), 54, 449
 - ERS-2—European remote-sensing satellite-2 (spacecraft), 26
 - ESTRACK. *See* European Space Tracking (ESTRACK) network
 - ETS-VII. *See* Engineering Test Satellite-VII (ETS-VII)
 - Euler
 - angles, 324
 - parameters, 325
 - rotational equations of motion, 327
 - European Cooperation for Space Standardization (ECSS), 31, 443
 - European Data Relay Satellite (ERDS), 7, 14, 29
 - European Space Tracking (ESTRACK) network, 29, 597
 - European X-ray observatory satellite (Exosat), 471
 - Eurostar (spacecraft) platform, 228, 350
 - Exosat. *See* European X-ray observatory satellite (Exosat)
 - Explorer-1 (spacecraft), 5
 - Extra-vehicular activity (EVA), 493
 - Extreme ultraviolet (EUV or XUV), 39
 - Extremophilic organism, 497
- F**
- Failure, 449, 628
 - blanching, 184
 - buckling, 209
 - corrosion fatigue, 237
 - detection, isolation and recovery (FDIR). *See* Fault, detection, isolation and recovery (FDIR)
 - dog-house effect, 184
 - fatigue, 184
 - hydrogen embrittlement, 206, 237
 - mode effects analysis (FMEA). *See* Hazard analyses
 - mode, effects, and criticality analysis (FMECA). *See* Hazard analyses
 - rupture, 184

F (cont.)

- stress corrosion, 237
 - stress corrosion cracking (SCC), 206
 - Fairing (launch vehicle). *See* Launch
 - Faraday cage, 52
 - Faraday rotation, 405
 - Fastening. *See* Joining
 - Fault
 - Avoidance, 449
 - correction mechanisms, 449
 - detection and isolation, 449
 - detection, isolation and recovery (FDIR), 351, 442, 449
 - management, 135
 - removal, 449
 - tolerance, 449
 - tree analyses (FTA), 55 *See also* Hazard analyses
 - FDIR. *See* Fault, detection, isolation and recovery (FDIR)
 - Fiber-reinforced polymer (FPR), 207
 - File delivery protocol (CFDP), 482
 - File-based operations, 581
 - Filter, 353
 - design, 354
 - fading memory, 366
 - Kalman. *See* Kalman filtering
 - Kalman-Schmidt, 366
 - least squares, 366
 - noise, 355
 - Finite burn losses, 94
 - Finite element analysis (FEA), 209, 212
 - Finite element method magnetics (FEMM), 237
 - Fire safety, 498
 - First point in aries. *See* Zero point of longitude
 - Flight dynamics system, 352, 582
 - Flight experiments, 109
 - Flight model (FM), 214
 - Flight operations segment, 28, 29, 575
 - human error, 576
 - link security, 586
 - mission control systems (MCS), 578
 - plan, 576
 - planning systems, 581
 - procedure (FOP), 576
 - automation, 577
 - classification, 577
 - contingency, 576
 - nominal, 576
 - test, 577
 - validation, 577
 - Flight software, 192, 471, 597
 - architecture, 479
 - boot software, 472, 480
 - command and data handling, 472, 480
 - commands, 481
 - compression applications, 484
 - design, 476
 - development, 474, 475
 - development models. *See* Systems engineering
 - embedded techniques, 477
 - external interfaces, 473
 - guidance, navigation, and control, 472
 - guidelines, 477
 - history, 471
 - image processing, 484
 - implementation, 477
 - in-chassis communication, 473
 - industry standards, 476
 - life cycle. *See* Life cycle
 - patch, 479
 - payload, 472, 485
 - planning, 475, 484
 - post launch, 479
 - requirements, 475
 - review, 476
 - simulator, 478
 - static analysis, 478
 - telemetry, 481
 - testing, 478
 - acceptance, 479
 - application, 478
 - integration, 478
 - operational, 479
 - scenario, 479
 - unit, 478
 - time and space partitioning, 477
 - updates, 479
 - Flight-path angle, 65, 379
 - Fluid loop, 386
 - Flywheel. *See* Reaction wheel
 - Food, 505
 - cooking, 506
 - nutrition requirement, 506
 - storage, 506
 - Formation flying, 96
 - definition of, 26
 - Fourier transform infrared spectrometer (FTIR), 494
 - Fourier's law, 110, 371
 - Fracture control, 212
 - Free radical, 497
 - Free-molecular (flow), 110
 - Frequency
 - allocation, 399
 - plan and channelization flexibility, 435
 - reuse, 424
 - spectrum, 398, 405
 - Frozen orbit, 690
 - Fuel cell, 271
 - electrochemistry, 272
 - performance, 272
 - Fuel mass fraction, 94
 - Functional
 - architecture, 151
 - decomposition, 150, 153
 - flow diagram, 151
 - n-square diagram, 151
 - timeline analysis, 151
- G**
- Gagarin, Yuri, 8, 11
 - Gaia (spacecraft), 17, 20, 222
 - Galactic cosmic ray (GCR), 42, 125, 496
 - Galactic radiation and background-1 (spacecraft). *See* GRAB-1 (spacecraft)
 - Galilei, Galileo, 75
 - Galileo (spacecraft), 37, 100, 520
 - flight software, 472
 - Galileo navigation system, 15
 - Gas chromatograph/mass spectrometer (GC/MS), 494
 - Gas regeneration, 499
 - Gaussian co-ordinate system. *See* Coordinate systems
 - Gegenschein, 50
 - Gemini (program), 101

- General relativity (GR). *See* Disturbing forces
- Generalized polynomial chaos expansions (GPCE). *See* Stochastic expansion methods
- Geocentric celestial co-ordinate system (GCRF). *See* Coordinate systems
- Geoid, 69
- Geometrical factor. *See* View factors, 378
- Geostationary transfer orbit (GTO), 89
- Geosynchronous orbit, 42, 89
geostationary orbit (GEO), 89
- Geotropism. *See* Gravitropism, 502
- Giotto (spacecraft), 548
- Glass, 222
Zerodur, 222
- Glenn, John, 8
- Glide trajectory, 102
- Global navigation satellite system (GNSS), 14, 18, 26, 344
- Global positioning system, 15, 26, 344
- GLONASS, 15, 18, 26
- Glow. *See* Atmospheric glow, 49
- Glushko, Valentin, 179
- GNC. *See* Guidance, navigation and control (GNC)
- GOCE. *See* Gravity field and steady-state ocean circulation explorer
- Goddard, Robert, 2, 179
- Google earth, 16
- Government-industry data exchange program (GIDEP) alerts, 52
- GRAB-1 (spacecraft), 5
- Graphene, 137
- Gravi-resistance reaction, 502
- Gravitational
constant, 62
parameter, 63
- Gravitropism, 502
- Gravity assist, 95
B-plane, 95
- Gravity field and steady-state ocean circulation explorer, 29, 30
- Gravity recovery and climate experiment, 16
- Gravity turn, 170, 531, 532
- Greenwich meridian, 67
- Ground control segment. *See* Flight operations segment
- Ground mission segment. *See* Payload data ground segment
- Ground segment, 25, 28, 30, 397, 401, 575
flight operations segment. *See* Flight operations segment
payload data ground segment. *See* Payload data ground segment
- Ground station, 29
- Ground track, 70, 690
- Grounding methods (electric), 55
- GSAT-3 (spacecraft). *See* EDUSAT (spacecraft)
- Guidance, navigation and control (GNC), 351
autonomous, 361
design, 352
flight software, 485. *See* Flight software
orbit control. *See* Orbit, control
orbit guidance. *See* Orbit, guidance
orbit navigation. *See* Orbit, navigation
rendezvous and docking, 361
- Guidance, navigation, and control software
algorithm development, 486
implementation, 487
- Gyro, 225. *See* Rate gyro
fiber optical (FOG), 225
hemispherical resonating (HRG), 225
- H**
- Harmonic analysis, 209
- Harness. *See* Power, distribution harness
- Hayabusa (spacecraft), 88, 101, 304, 345, 367, 548
- Hazard analyses, 55, 159
fault-forecasting, 449
- HDRM. *See* Mechanisms
- Health and safety, 135, 170
- Heat flux density, 372
- Heat pipe, 384
- Heat switch, 386
- Heat transfer, 371
catalytic reactions, 106
conductive, 371
radiative, 106, 108
- Heater, 386
- Heilmeyer questions, 608
- Heilmeyer, George, 608
- Heisenberg uncertainty principle, 626
- HERMES (communications spacecraft). *See* Communications technology satellite ('Hermes')
- Hermes (spaceplane), 203
- Herschel (spacecraft), 198, 212, 215, 222
- Highly elliptical orbit (HEO), 90
- High-order assembly language/shuttle (HAL/S). *See* Programming language
- Hikoboshi (spacecraft). *See* Engineering Test Satellite-VII (ETS-VII)
- Hill, George William, 97
- Hill's equations. *See* Clohessy-Wiltshire equations of motion
- Hills limiting surface, 74
- Hipparcos (spacecraft), 17
- Hipparcos star catalog, 17, 342
- Hocquenghem, Alexis, 409
- Hohmann transfer, 92
- Hohmann, Walter, 3
- Hold-down and release mechanisms (HDRM). *See* Mechanisms
- Horizon sensor, 340
- Housekeeping data, 481
- HST. *See* Hubble space telescope (spacecraft)
- Hubble space telescope (spacecraft), 6, 17, 45, 219, 241
- Human space flight, 7, 11
- Huygens probe (spacecraft). *See* Cassini-Huygens (mission)
- Hybrid propulsion development program, 300
- Hybrid sounding rocket (HYSR), 300
- Hydrodynamic equations, 113
- Hylas-1 (spacecraft), 436
- Hyperoxic, 493
- Hypersonic, 103
- Hyperstability, 229
- Hypervelocity impacts, 49
- Hypoxia, 493
- I**
- IFOC. *See* Rocket motor, igniter
- Ignition overpressure (IOP), 175
- IKAROS (spacecraft), 318
- Impulse, 279
specific, 94, 167, 279
- Inclination, 66
- Inertial measurement units (IMU), 344
- Inertial navigation system (INS), 356
- Inertial reference unit (IRU), 344
- Inertial stellar compass (ISC), 337

I (cont.)

Infrared (IR), 40
 Innovation, 599, 608, 610, 617
 Input filter assembly, 425
 Instrument. *See* Payload
 Integrated in-space transportation plan, 681
 Intellectual property, 608, 617
 Intelsat
 -1, 5, 397
 -708, 35
 Interface
 control document (ICD), 123, 135, 146
 requirements documents (IRD), 146
 Internal electrostatic charging/discharging (IESD), 47
 International Celestial Reference System (ICRS). *See* Coordinate systems
 International Commission on Radiological Protection (ICRP), 497
 International space station, 10, 11, 101, 303, 346
 alpha magnetic spectrometer, 11
 columbus laboratory, 200
 Crew Return Vehicle (CRV), 203, 522
 cupola, 199
 Remote Manipulator System (SSRMS), 543
 SOLAR payload, 201
 International Traffic in Arms Regulations (ITAR), 35
 1247 report, 36
 2013 defense authorization bill, 36
 arms export control act, 35
 United States Munitions List (USML), 35
 Interplanetary Kite-craft accelerated by radiation of the sun (spacecraft). *See* IKAROS (spacecraft)
 Inter-satellite link, 26
 Inverse Cheng parameter, 110
 Ionization, 109
 Iridium (spacecraft), 26
 IRNSS (navigation system), 15
 Irradiation. *See* Total emissive power
 Isolator, 218
 Iso-static, 212
 ITAR. *See* International Traffic in Arms Regulations (ITAR)
 Itokawa asteroid, 304, 548
 ITU-R model. *See* Simulation, rain attenuation

J

Jacobi
 constant, 73
 generalized Jacobian, 552
 integral, 73
 James Webb Space Telescope (JWST), 198, 240
 Japanese Experiment Module Remote Manipulator System (JEMRMS), 543
 Jet power, 281
 Joan of Arc, 187
 Joining (process), 225
 Joint (thermal) resistance. *See* Thermal contact
 Julian
 date, 72
 period, 72
 Juno (spacecraft), 48, 55
 Juste retour, 166

K

Kalman filtering, 354, 365
 extended, 331

Kalman-Schmidt, 367
 stability, 354
 unscented (UKF), 332
 Kármán line. *See* von Kármán ellipsoid
 Kaufmann, Harold, 303
 Kelly cosine, 253, 340
 Kennedy, John F., 8
 Kepler (spacecraft), 20, 341, 345
 Kepler, Johannes, 61
 Kepler's
 equation, 65
 laws, 61
 Keplerian
 motion, 61
 orbital elements. *See* Classical orbital elements
 Kevlar, 207
 Key decision points (KDP). *See* Schedule, milestone
 Kinematic equations, 326, 550
 Kirchhoff's law, 378
 Knudsen number, 85, 105, 110, 520
 Komarov, Vladimir, 9
 Korolev, Sergei, 2, 4, 179, 298
 Korolyov, Sergey. *See* Korolev, Sergei
 Kuchemann's correlation, 103

L

Lagrancia, Giuseppe Luigi. *See* Lagrange, Joseph-Louis
 Lagrange points, 74
 Lagrange, Joseph-Louis, 74, 77, 78
 Lagrange's planetary equations, 77
 Laika (dog), 8
 Lambert's problem, 91, 691
 Laminar, 105
 laminar-turbulent transition, 105
 Landing, 516, 532
 airbag. *See* Airbag
 definition of, 515
 legs, 533
 mid-air retrieval, 534
 Lang, Friedrich Christian Anton 'Fritz', 3
 Laser Interferometer Space Antenna (LISA) mission, 27
 Latchup (single event effect), 45
 Latitude, 67
 Launch
 control, 172
 fairing, 209
 guidance, 172
 navigation, 172
 profile, 169
 trajectory, 170
 vehicle, 26, 165, 202
 Launch and Early Orbit Phase (LEOP), 591
 Lavochnik, Semyon, 533
 Le Verrier, Urbain, 75
 Legal. *See* Space law
 Legendre polynomial, 81
 Lewis, Gilbert N., 86, 318
 Life cycle, 32, 149, 157, 473, 599, 605, 623, 624, 631, 636, 692
 flight operations, 590
 model. *See* Systems engineering
 phase, 149
 software, 157
 technology, 599
 whale chart, 599
 Life management, 505

- Life support, 493, 499
 - controlled ecological life support system (CELSS), 501
 - Lift coefficient, 102, 167, 174
 - Lift-to-drag ratio, 103
 - Lightcurve, 367
 - Linear energy transfer (LET), 45
 - Linear momentum, 551
 - Link (communications)
 - down, 401
 - forward, 401
 - inter-satellite, 401, 404
 - return, 401
 - up, 401
 - LISA Pathfinder (spacecraft), 225, 340
 - Lǐwěi, Yáng, 11
 - Load events, 210
 - Load levels
 - acceptance, 177
 - acceptance factor, 203
 - flight limit loads (FLL), 210
 - limit load, 177
 - proof. *See* Load levels, acceptance
 - quasi-static loads (QSL), 208
 - safety factor, 176, 177, 203, 210, 217
 - test load factors, 210
 - ultimate, 177
 - yield, 177
 - Localization (rover), 562
 - Long Duration Exposure Facility (LDEF), 48
 - Longitude, 67
 - Lorentz effect, 48
 - Louvers, 383
 - Low earth orbit (LEO), 89
 - Low voltage differential signaling (LVDS), 462
 - Lubricant, 235
 - Luna-2 (spacecraft), 5
 - Lunar surveyor (spacecraft), 516, 531
 - Lunokhod (spacecraft), 545
 - Lyman-alpha line, 40
 - Lynx (spaceplane), 16, 671
- M**
- Magellan (spacecraft), 96
 - flight software, 472
 - Magnetic field, 39, 107
 - Magnetometer, 336, 343
 - Magnetorquers, 346
 - MagOrion project, 317
 - Management information systems (MIS), 646
 - Mango and Tango (spacecraft). *See* Prisma (spacecraft)
 - Manmade debris. *See* Synthetic debris
 - Manufacturing, 222
 - additive manufacturing, 224
 - autofrettage, 223
 - casting, 223
 - fiber steering, 224
 - forging, 223
 - forming, 222
 - machining, 222
 - resin transfer molding, 224
 - Margin. *See* Contingency
 - Margins of safety (MS). *See* Load levels, safety factor
 - Mariner probe
 - 10, 88, 95, 344
 - Flight software, 471
 - Mars exploration rover, 7, 100, 526, 532, 546, 548
 - Mars global surveyor (spacecraft), 96
 - Mars pathfinder (spacecraft), 100, 517, 532, 546, 547
 - Mars reconnaissance orbiter, 352
 - Mars science laboratory (spacecraft), 100, 229, 356, 547
 - descent propulsion system, 531
 - sky-crane, 532, 547
 - Mars Sojourner rover. *See* Mars pathfinder
 - Marskhod (spacecraft), 546
 - Mass memory, 444, 453
 - architecture, 453
 - Maximum expected operating pressure (MEOP), 177
 - Maxwell, James Clerk, 86
 - Mean anomaly, 65
 - Mean motion, 65
 - Mean time between failures (MTBF), 166
 - Mean-free-path, 110
 - Mechanisms, 225
 - deployment mechanisms, 230
 - hold-down and release mechanisms (HDRM), 230
 - materials, 237
 - pointing
 - antenna pointing mechanisms (APMs), 229
 - coarse pointing assembly (CPA), 229
 - electric propulsion pointing mechanisms (EPPMs), 228
 - solar array drive mechanism (SADM), 233
 - position sensors, 232
 - optical encoder, 232
 - potentiometer, 232
 - switch, 232
 - verification. *See* Verification & validation
 - Medical support, 509
 - medication, 511
 - Medium Earth Orbit (MEO), 89
 - Méliès, Georges, 3
 - Meosat-3b (spacecraft), 350
 - Mercury (project), 101
 - Mercury Sigma-7 (spacecraft), 30
 - Mercury Surface, Space ENvironment, GEochemistry and Ranging. *See* MESSENGER
 - MESSENGER (spacecraft), 7, 39, 88, 99, 344
 - flight software, 488
 - spacecraft overview, 488
 - Metabolic rate, 506
 - Meteoroids, 44, 125
 - Meteorology, 12
 - Micrometeoroids. *See* Meteoroids
 - Microwave instrument, 452
 - Military space, 18
 - Military surveillance, 19
 - MIL-STD-1553, 456, 483
 - bus controller (BC), 456, 457
 - protocols, 456
 - scheduling, 457
 - Minimum impulse bit (MIB), 359
 - Minimum mission. *See* Threshold Mission
 - Mir (spacecraft), 9, 101
 - MISRA-C. *See* Flight software, guidelines
 - Missile defense, 20
 - Missile gap, 8
 - Mission analysis, 687
 - Molecular free path, 105
 - Molniya
 - orbit, 91
 - spacecraft, 5
 - Momentum wheel (MW). *See* Reaction wheel (RW)

M (*cont.*)

Monolithic passive flexible element (MEDI), 219
 Monopropellant. *See* Propellant
 Monte-Carlo analysis, 171, 538
 Moonlight, 51
 Moore, William, 2
 Moore's law, 454
 Mortar. *See* Pyrotechnic
 Motor Industry Software Reliability Association (MISRA) C. *See*
 Flight software, guidelines
 Multilayer insulations (MLI), 381
 second surface mirror, 382
 Multimedia internet access, 401
 Multipacting, 51
 Multiplexing, 409
 input multiplexer (IMUX), 412, 428
 output multiplexer (OMUX), 430
 Multi-stage (rocket), 169

N

Navier–stokes (equations), 109, 110, 112, 282, 301
 Navier-Stokes-Fourier (equations). *See* Navier-Stokes (equations)
 Near Earth Network (NEN), 29
 NEAR-Shoemaker, 367, 548
 NERVA (project), 314
 Network. *See* Data routing
 Neutral axis maneuver, 172
 New millennium program ST-6, 337
 Newton, Isaac, 62
 Newton's law
 of motion, 62
 of universal gravitation, 62
 Newtonian methods, 112
 NigeriaSat-2 (spacecraft), 216
 Nobel, Alfred, 187
 Noise power density, 403
 Nonholonomic path planning, 554
 Non-processing payload. *See* Communication system
 Noordung, Hermann. *See* Potočnik, Herman
 Nozzle. *See* Rocket motor or rocket engine
 NSF. *See* Navier-Stokes (equations)
 Nuclear Engine for Rocket Vehicle Application (NERVA). *See*
 NERVA (project)
 Nuclear safety launch approval, 55
 Numerical integration, 80
 errors, 80
 Nusselt relation, 183

O

Oberth, Hermann, 3, 179
 Obliquity of the ecliptic, 67
 Oil flares. *See* Stray light
 On-board software (OBSW). *See* Flight software
 On-orbit servicing, 542
 Operational procedures, 55
 Operations concept, 133
 Operations simulator, 30, 577, 583, 591
 architecture, 583
 benefits, 585
 campaign, 585
 Opportunity management, 643
 Opportunity rover. *See* Mars exploration rover
 Optical imaging instrument, 451
 Orbit

control, 359, 690
 thruster management, 359
 definition of, 61
 design, 688
 determination, 582
 dynamics, 61
 energy, 65
 guidance, 356
 nominal trajectory, 357
 predictive-impulsive, 366
 maintenance, 594
 navigation, 353
 measurement types, 355
 proportional navigation, 366
 vision-based, 356, 362
 period, 65
 selection, 127
 stationkeeping, 691
 velocity, 65

Orbital debris. *See* Synthetic debris or meteoroids

Orbital energy conservation equation. *See* Vis-viva equation

Orbital express (spacecraft), 544

Orihime (spacecraft). *See* Engineering Test Satellite-VII (ETS-VII)

Outer space

definition of, 657s

exploitation, 669

Outgassing, 208

Oxygen

candle, 500

erosion, 48, 125, 135

P

Palapa (spacecraft), 398

Parachute, 104, 204, 516, 522

conical ribbon, 528

cross. *See* Parachute, cruciform

cruciform, 529

deployment, 530

disk-gap-band (DGB), 527

inflation, 527

mass ratio, 524

parafoil, 530

polyconical, 530

porosity, 523

ringsail, 529

ringslot, 528

Parasitic power losses. *See* Power losses

Pascal, Blaise, 187

Paschen's law, 262

Path planning (rover), 564

Payload

accommodation, 117, 122

attitude control, 134

calibration, 134

categories, 119

classification, 120

command and control, 134

communications. *See* Communication system

concept design, 118, 128, 131, 132

cryogenic, 139

data system. *See* Data system

definition of, 27, 117

design process, 118

environment, 122, 124

ground data segment. *See* Payload data ground segment

- integration, 123
- interfaces, 123
- operation, 133
- performance, 130
- software, 472, 485
- strawman, 131
- thermal, 122, 129
- verification, 127
- Payload data ground segment, 28, 587
 - architecture, 587
 - auxiliary data files (ADF), 587
 - data levels, 587
 - user segment, 589
- Payload data levels. *See* Payload data ground segment
- PD Controller. *See* Controller, proportional-derivative (PD)
- Pebble bed reactors (PBR), 315
- Pegasus (rocket), 169
- Peltier effect, 387
- Periapsis, 64
- Pericentre. *See* Periapsis
- Perigee. *See* Periapsis
- Perihelion. *See* Periapsis
- Peroxide, 497
- Perturbation techniques, 75
 - Cowell's method, 76
 - Encke's method, 76
 - general perturbations, 75
 - special perturbations, 75
 - variation of parameters, 77
- Perturbations. *See* Disturbing forces
- Phase-change material (PCM), 385
- Phase noise spectral density, 427
- Phobos (mission), 546
- Phoenix lander (spacecraft), 100, 546
 - descent propulsion system, 531
- Photoelectron emission, 40
- Photon, 86, 318
- Photoperiodism, 502
- Photosynthesis, 502
- Photovoltaic (PV)
 - cell, 250
 - inverted metamorphic multijunction (IMM), 250
 - margin, 253
 - thin-film, 250
- Photovoltaic-battery system
 - architectures, 250
 - peak power tracking (PPT), 250
 - regulated bus, 250
- Phugoid, 104
- Physiology, 510
- PID controller. *See* Controller, proportional-integral-derivative (PID)
- Pioneer Venus (spacecraft), 100, 520
- Pitching moment anomaly, 111
- Planck (spacecraft), 140, 198
- Planck's
 - constant, 376
 - law, 86, 318
- Planetary protection, 505, 667
- Planetary radiation, 40, 380
 - equivalent black-body temperature, 380
- Planetary ring. *See* Meteoroids
- Plasmasheath, 46
- Plasma sheet, 42
- PMBOK. *See* Project management institute body of knowledge
- Pogo oscillations. *See* Rocket engine, combustion instabilities
- Pontryagin principle, 171
- Potočník, Herman, 397
- Power
 - advanced technology, 137
 - battery. *See* Battery
 - beamed power, 273
 - brayton cycle, 270, 291
 - bus Impedance, 263
 - Carnot cycle efficiency, 269, 270
 - deep space mission, 267
 - distribution harness, 259
 - dynamic system with alternator, 270
 - electronics, 257
 - shunt regulator, 259
 - switching devices, 257
 - energy balance, 263
 - fuel cell. *See* Fuel cell
 - interplanetary mission, 266
 - management, 263
 - near-sun mission, 267
 - reliability, 266
 - solar array. *See* Solar array
 - stirling radioisotope generator (SRG), 137
 - system design, 261
 - system performance, 263
 - system requirements, 262
 - system software, 473, 485
 - system stability, 263, 264
 - tetrapods, 137
 - transfer, 123
 - ultracapacitors, 137
- Power losses, 48
- PowerPC, 455
- Prandtl number, 105
- Pressure vessel, 177, 202, 206
 - composite overwrapped (COPV), 177, 202, 207
 - minimum burst factors, 207
 - pressurized component, 207
 - testing, 207
 - ultimate strength, 207
- Pressurized component. *See* Pressure vessel, pressurized component
- Pressurized system. *See* Pressure vessel
- Primer. *See* Rocket motor, igniter
- PRINCE. *See* PRojects IN Controlled Environments
- Prisma (spacecraft), 298, 352
- Process data objects (PDOs), 460
- Processor, 473
 - benchmarking, 473
 - LEON, 455
 - microprocessor without Interlocked Pipeline Stages (MIPS)
 - architecture, 455
 - PowerPC, 455
 - RAD, 455
 - RHPPC, 455
 - SPARC, 455
- Product assurance, 166
- Production constraints, 166
- Productization, 601, 603
- Prograde, 66, 70
 - apparent, 70
- Programming language, 471, 472, 475, 477
- Project management, 619
 - framework, 623
 - maturity (PMM), 624
 - overview, 622
 - practice, 626
 - strategies, 624

P (*cont.*)

Project Management Institute Body of Knowledge (PMBOK), **623**

Project Orion, **316**

Propellant

- bipropellant, **177, 289, 295**
- ceramic fuel (CERMET), **315**
- cold gas, **177**
- liquids, **179, 295**
- management, **594**
- monopropellant, **177, 288, 297**
- propellant grain, **189, 283, 287**
- propellant slag, **283**
- solid propellant, **178, 283**
- tank, **176, 206**
- tripropellant, **291**

Propulsion, **279, 548, 688**. *See* Rocket engine or motor

- advanced, **313**
- advanced in-space propulsion (ISP) technologies, **681**
- attitude control, **348**
- electric, **138, 228, 281, 300, 315, 548**
- electrostatic, **303**
- electrothermal, **306**
- Magnetoplasmadynamic (MPD), **309**
- power, **303, 305, 308**
- propellant, **302, 305, 312**
- liquid propulsion, **178, 280, 288**
- minimum impulse bit (MIB), **359**
- nuclear, **314**
- saturation, **359**
- solar sail, **138, 240, 314, 317, 594**
- solid propellant, **187, 281**

Proto flight model (PFM), **214**

Prototype, **602**

Pyrotechnic, **231**

- igniter. *See* Rocket motor, igniter
- mortar, **530**

Q

Quaternion, **325**

QZSS (navigation system), **15**

R

Radiation belts, **42**

Radiation models, **42**

- AE8/AP8 model, **42**
- AE9/AP9 model, **42**
- GIRE, **42**
- SATRAD, **42**

Radiative losses, **308**

Radiator, **382**

- active, **382**
- efficiency, **383**
- passive, **382**

Radio frequency (RF) propagation, **50**

Radioisotope thermoelectric generator (RTG), **55, 267, 268, 386**

- half-life, **270**
- seebeck effect, **269**

Radiosity. *See* Total emissive power

Range finder, **355**

- laser, **367**

Rarefaction, **109, 113**

- bridging functions, **110**
- slip regime, **110**

Rarefied

flow, **105, 110–112**

gas, **106**

Rate gyro, **344**

Reaction wheel (RW), **225, 337, 345**

Real-gas effects, **109, 113**

Real-time operating systems, **477**

Receive antenna. *See* Antenna

Receiver, **413, 426**

Recovered mass loss (RML), **208**

Rectilinear orbit, **65**

Recycling, **499**

physicochemical system, **499**

water, **499**

Redundancy, **52**

Reefing, **527**

Reentry vehicle. *See* Atmospheric entry vehicle, **203**

Reference frame. *See* Coordinate systems

Reflectance

spectral directional, **377**

spectral hemispherical, **378**

Regenerative cooling, **292**

Relative biological effectiveness (RBE), **497**

Reliability analysis, **706**

Remote memory access protocol (RMAP), **465**

Remote terminal, **456**

Requirements, **143, 150, 616, 698**

allocated, **153**

commercial payloads, **121, 131**

constraints, **135, 352**

data management, **127, 129, 132**

derived, **153**

descope, **641**

generation, **120**

materials, **219**

measurement, **132**

mechanical, **208**

military payloads, **121, 131**

mission resources, **128**

needs assessment, **149**

operational, **121, 131**

orbit, **123, 135**

performance, **130**

power system, **262**

science payloads, **120, 131, 136**

spacecraft, **134**

system-level, **150**

traceability, **153**

traceability matrix, **122, 153**

Research and development degree of difficulty. *See* Advancement degree of difficulty, **34**

Resistive bleed path. *See* Grounding methods, **55**

Resistivity

sheet, **266**

Retrograde, **66, 70**

apparent, **70**

Reverberant acoustic room, **215**

Review

action items, **154**

flight software. *See* Flight software

panel, **153**

technical assessment, **153**

Reynolds

analogy, **106**

number, **105, 110, 174, 520**

Right ascension of the ascending node, **66**

Risk, **607, 617**

- analysis, 703
- categorization, 159
- classification, 160
- control, 159
- management, 159, 641, 705
- matrix, 159
- planning, 159
- posture, 642
- reward, 606
- tracking, 159
- Riveting. *See* Joining
- Robonaut, 545
- Robotics, 541, 545
 - control, 549
 - dynamics, 558
 - human interface, 567
 - kinematics, 556, 557
 - simulation, 557
 - trajectory, 554
- Robot technology experiment (ROTEX), 544
- Rocket engine
 - combustion chamber, 182
 - combustion instabilities, 184, 291, 295, 300
 - cooling, 293
 - cryogenic, 289
 - cycle, 180, 291
 - feed, 179
 - gas generator, 185
 - hybrid, 179, 298
 - ignition system, 182
 - injection system, 180
 - nozzle, 185
 - non-equilibrium effects, 295
 - pre-burner, 185
 - re-ignition, 294, 298
 - thermodynamic rockets, 282
 - throttle, 295
 - turbines, 186
 - turbopump, 186, 289
- Rocket equation, 94, 168, 280
- Rocket motor, 187
 - combustion instabilities, 287
 - igniter, 191
 - motor case, 188
 - nozzle, 191
 - pressure oscillations, 192
 - propellant grain, 189, 283
 - propellant slag, 283
 - segmented, 188
 - thermal protection, 189
- Rocket propulsion. *See* Propulsion
- Rosetta (spacecraft), 549
- Rotational kinematics, 323
- Rover (project), 314

- S**
- Sabatier processor, 499
- Safety constraints. *See* Health and safety
- Salyut (spacecraft), 9, 101
- Sample-return (mission), 548
- Sandwich panels, 197
- Satcom (spacecraft), 398
- Satellite
 - definition of, 25
- Saturated power flux density (SFD), 418
- Schedule, 160, 636, 704
 - critical path, 160, 636
 - earned value schedule-performance-index (SPI), 636
 - estimate of schedule at completion (ESAC), 162
 - float, 160, 636
 - gantt chart, 160
 - integrated mission schedule (IMS), 636
 - key decision point (KDP), 623
 - milestone, 160, 636
 - network diagrams, 160
 - percent complete (PC), 161
 - schedule at completion (SAC), 161
 - schedule performance index (SPI), 161
 - schedule variance (SV), 161
- SCORE (project), 5
- Search and rescue, 13
- Sectoral harmonic. *See* Disturbing forces, non-spherical central body
- Seebeck effect. *See* Radioisotope thermoelectric generator (RTG)
- Semi-latus rectum, 65
- Semi-major axis, 64
- Semi-minor axis, 64
- Sensitivity analysis. *See* Trade study, sensitivity analysis
- Service data objects (SDOs), 460
- Shafer's evidence theory (ET), 710
- Shaker table, 214, 215
- Shannon limit, 409
- Shannon's information theorem, 403
- Shape memory alloys, 232
- Shenzhou (spacecraft), 11
- Shepard, Alan, 8
- Shielding, 52
 - doghouse, 52
 - vault, 52
- Shock, 105, 109
 - layer, 106
 - response spectrum (SRS), 209, 212, 231
- Shuttle. *See* Space shuttle
- Signal attenuation
 - atmospheric, 403, 405
 - Ionospheric effects, 405
 - rain, 404, 406
- Signal Communications Orbit Relay Equipment. *See* SCORE (project)
- Signal error correction, 586
 - adaptive coding and modulation (ACM), 409
 - automatic repeat request (ARQ), 408
 - coding, 408, 445, 453
 - forward error correction (FEC), 408, 586
- Signal interference, 410
- Signal modulation
 - analog, 406
 - asymmetric phase-shift keying (APSK), 407
 - bandwidth efficiency, 407
 - constant envelope modulation (CEM), 407
 - digital, 406
 - M-ary frequency shift keying (MFSK), 407
 - M-ary phase shift keying (MPSK), 407
 - M-ary quadrature amplitude modulation (MQAM), 407
 - power efficiency, 407
- Signal-to-noise ratio, 132, 134, 402
- Silicon carbide (SiC). *See* Ceramics

S (cont.)

- Silicon controlled rectifier (SCR). *See* Latchup
- Silicon-on-insulator (SOI), 455
- Silkworm, 504
- Simulation
 - aerodynamics, 112
 - deployable structures, 240
 - Direct-Simulation Monte Carlo (DSMC), 110, 111, 113
 - electromagnetic, 237
 - entry, descent and landing, 534
 - multi-body dynamic, 236
 - multi-physics, 237
 - operations. *See* operations simulator
 - rain attenuation, 406
 - robotics, 556
 - test simulators. *See* flight software, simulators
 - thermal model, 374, 391
 - wheel contact model, 559
- Single event effects. *See* Upsets (SEU)
- Single-stage to orbit, 114, 169, 195
- Skip trajectory, 102
- Sky-crane. *See* Mars Science Laboratory (spacecraft)
- Skylab (spacecraft), 9, 101, 619
- Skylon (spaceplane), 114
- Sloshing. *See* Disturbing forces, internal torques
- SMART, specific, measurable, attainable, relevant, and timely, 614
- SMART-1 (spacecraft), 228, 312
- Software. *See* Flight software
- Software-defined radios, 485
- SOHO (spacecraft), 134
- Solar and heliospheric observatory (spacecraft). *See* SOHO (spacecraft)
- Solar array
 - degradation, 255
 - drive electronics (SAD/ADE), 260
 - drive mechanism (SADM). *See* Mechanisms, pointing
 - electro-explosive deployment (EED), 261
 - peak power tracking, 255
 - performance, 253
 - Photovoltaic (PV). *See* Photovoltaic (PV)
 - P-V and I-V characteristics, 252
- Solar constant. *See* Solar spectrum
- Solar energetic particle (SEP) event, 43
- Solar flare, 43
- Solar irradiance. *See* Solar spectrum
- SOLAR payload. *See* International space station
- Solar proton event (SPE), 43
- Solar radiation. *See* Solar spectrum
- Solar spectrum, 40, 379
 - ASTM E490, 40
 - black-body, 40, 377
 - irradiance, 40, 87, 379
 - ISO-21348, 40
 - zero air mass solar spectral irradiance, 40
- Solar wind, 40, 88, 107
 - L2-CPE statistical model, 40
 - Nascap-2K, 41
- Soldering. *See* Joining
- Solid propellant. *See* Propellant
- Solid rocket motor (SRM). *See* Propulsion or propellant
- Sound pressure levels (SPL), 215
- South Atlantic Anomaly (SAA), 125, 128, 135
- Soyuz, 9
 - capsule, 9, 203
 - descent module. *See* Soyuz, capsule
 - launcher, 4, 9
- Space 2.0, 7
- Space age, 1, 5
- Space architecture, 508
- Space-based infrared system, 20
- Spacecraft
 - commissioning, 592
 - definition of, 25
 - operations. *See* ground segment
 - operator, 575
- Space debris. *See* Synthetic debris or meteoroids
- Space Environment Information System (SPENVIS), 58
- Space flight participants, 16
- Space law, 657
 - liability convention, 661
 - moon agreement, 666
 - national, 663
 - outer space treaty, 658
 - registration convention, 662
 - rescue agreement, 660
- Space motion sickness (SMS), 510
- Space network (SN), 29
- Space probe
 - definition of, 25
- Space race, 5
- Space science, 16
- Space segment, 25, 26, 397, 401
- SpaceShipOne, 101, 298, 300
- SpaceShipTwo, 16, 114, 671
- Space shuttle, 9, 101, 105, 111, 515, 517, 520
 - flight software, 471
 - main engine (SSME), 179
 - radar topography mission (SRTM), 241
 - Remote Manipulator System (SRMS), 542
 - tether experiment, 55
- Space system, 143
 - definition of, 25
- Space tourism, 16, 671
- Space vehicle
 - definition of, 25
- Spaceway, 400
- SpaceWire, 461, 484
 - architecture, 464
 - links, 462
 - network, 463
 - remote memory access protocol (RMAP), 465
 - time-codes, 465
- Special Purpose Dexterous Manipulator (SPDM), 543
- Special relativity, 86
- Sphere of influence (gravity), 80
- Spin-in technology. *See* Technology infusion
- Spin-off technology. *See* Technology infusion
- Spiral development model. *See* Systems engineering, spiral development model
- Spirit rover. *See* Mars exploration rover
- Spitzer (spacecraft), 129
- Spot beam, 424
- Sputnik (spacecraft)
 - 2, 8
 - 3, 8
 - 4. *See* Vostok (spacecraft)
 - 5. *See* Vostok (spacecraft)
 - crisis, 1
 - Korabl-Sputnik-1. *See* Vostok (spacecraft)
 - Korabl-Sputnik-2. *See* Vostok (spacecraft)
 - program, 7
 - Sputnik-1, 1, 5

- Squib. *See* Rocket motor, igniter
 SSTO. *See* Single-stage to orbit
 Standard model, 280
 Stanton number, 110
 Star sensor, 341
 Stardust (spacecraft), 101, 104, 111, 365, 520, 548
 State vector, 66, 166
 Statistical energy analysis (SEA), 209
 Stefan-Boltzmann
 constant, 377
 law, 377
 Stentor (spacecraft), 350
 STEREO (spacecraft), 17
 Stereo vision, 562
 Stewart platform, 219
 Stochastic collocation (SC). *See* Stochastic expansion methods
 Stochastic expansion methods, 710
 Strawman payload. *See* Payload
 Stray light, 50
 Structural analyses
 acoustic, 212
 buckling, 212
 crack growth, 212
 dynamic displacement, 212
 fatigue growth
 frequency response, 212
 micro-vibration, 212
 modal, 212
 random response, 212
 stability, 177, 209
 static stress, 212
 transient, 212
 Structural ratio, 169
 Structural sizing
 stability, 177
 stiffness, 177
 strength, 177
 Structural thermal models (STM), 214
 Structure
 crewed, 199
 primary, 197
 secondary, 197
 Stuhlinger, Ernst, 282, 303
 Sun presence detector. *See* Sun sensor
 Sun sensor, 339
 Sunjammer (spacecraft), 318
 Sun-synchronous orbit, 90, 689
 Supersonic/hypersonic arbitrary body programs (SHABP), 112
 Surface charging, 46, 124, 125, 139
 Sutton and Graves correlation, 519
 Swarm (spacecraft), 215
 Syncom (spacecraft), 5, 397
 Synodic system. *See* Coordinate systems
 Synthetic aperture radar (SAR), 26, 452
 Synthetic debris, 44, 126
 DELTA (Debris environment long-term analysis) tool, 44
 graveyard maneuver, 595
 Inter-Agency Space Debris Coordination Committee, 89
 MASTER (Meteoroid and Space Debris Terrestrial Environment Reference) model, 44
 orbital debris engineering model ORDEM206, 44
 ORDEM2000, 44
 PROOF (Program for Radar and Observation Forecasting) model, 44
 System design. *See* System
 Systems engineer, 144
 dilemma, 146
 Systems engineering, 143
 agile methods model, 474
 definition of, 143
 incremental build model, 474
 lessons learned, 637, 652
 management plan (SEMP), 143, 144
 organization, 158
 rapid prototype model, 624
 software, 157, 473
 spiral development model, 145
 spiral model, 474, 626
 waterfall model, 145, 474
T
 TacSat-2 (spacecraft), 337
 Taikonaut, 11
 Tandem-X (spacecraft), 16
 Tauber-Sutton radiative heating correlation, 519
 TDRSS. *See* Tracking and Data Relay Satellite System
 Technology infusion, 20, 599, 603, 607, 616
 Technology manager, 614
 Technology Readiness Level (TRL), 33, 147, 599, 620
 relevant environment, 615
 valley of death, 602, 603, 614, 680
 Technology roadmap, 605
 Telecommand, 442, 481, 578
 database, 578
 execution verification, 580
 packet, 579
 parameter, 579
 pre-transmission validation (PTV), 579
 time tagging, 578, 581
 validation, 578
 Telemedicine, 512
 Telemetry, 444, 578
 calibration curve, 579
 database, 578
 derived parameters, 579
 Out of Limits (OOL) condition, 579
 packet, 579
 parameter, 579
 Telemetry, commands, and ranging (TC&R). *See* Telemetry, tracking, and commands (TT&C)
 Telemetry, tracking, and commands (TT&C), 415
 Teleoperation. *See* Telerobotics
 Telerobotics, 564
 Telstar (spacecraft), 5, 397
 Terminal velocity, 515
 Terra (spacecraft), 16
 Terrain mapping, 562
 Terramechanics, 559
 TerraSAR-X (spacecraft), 16
 TerreStar-1 (spacecraft), 422
 Tesser harmonic. *See* Disturbing forces, non-spherical central body
 Test plan. *See* Verification and validation, test plan
 Thermal
 conductivity, 372
 diffusivity, 372
 effects, 49
 management, 595
 radiation heat transfer, 375
 resistance, 372
 Thermal blanket. *See* Multilayer insulations
 Thermal contact

- T** (*cont.*)
- conductance, 384
 - resistance, 384
- Thermal control surface. *See* Control surface
- Thermal Control System (TCS), 389
- Thermal filler, 384
- Thermal protection system, 101, 105, 109, 204, 388, 518, 519
- ablative systems, 388
 - radiative systems, 388
 - transpiration systems, 388
- Thermodynamic efficiency, 291
- Thermoelectric cooling, 387
- Thermoluminescent dosimeter (TLD), 497
- Thevenin equivalent, 263
- Three-body problem, 73
- circular restricted (CRTBP), 73
 - elliptic restricted (ERTBP), 73
- Threshold mission, 131
- Thrust coefficient, 178, 286
- Thrust vector control (TVC), 233
- Thruster misalignment. *See* Disturbing forces, internal torques
- Tikhonravov, Mikhail, 298
- Time, 71
- Coordinated Universal Time (UTC), 72
 - Greenwich Mean Time (GMT), 72
 - International Atomic Time (TAI), 72
 - leap seconds, 72
 - mean solar day, 72
 - mean solar time, 72
 - Mission Elapsed Time (MET), 448
 - on-board (OBT), 447, 487
 - one-way light time (OWLT), 596
 - second, 72
 - sidereal day, 72
 - synchronization, 578, 580
 - Universal Time (UT), 72
- Ting, Samuel. *See* International Space Station, Alpha Magnetic Spectrometer
- TIROS-1 (spacecraft), 5, 12
- Tisserand, François Félix, 81
- Titius-Bode law. *See* Bode's law
- Tito, Dennis, 672
- Tokamak confinement, 305
- TopSat (spacecraft), 19
- Torques, 49
- Total dose. *See* Dose
- Total emissive power, 376
- Total ionizing dose (TID). *See* Dose
- Total organic carbon (TOC), 494
- TPS. *See* Thermal protection system
- Tracking and data relay satellite system, 5, 29
- Tractor rocket, 530
- Trade study, 127, 146
- sensitivity analysis, 147
- Traffic encryption key (TEK), 586
- Trajectory
- definition of, 61
 - selection, 55
- Transmittance
- spectral directional, 377
 - spectral hemispherical, 378
- Transponder, 400
- Traveling wave tube amplifiers (TWTA), 412, 414, 432, 436
- Treaty of Versailles, 3
- Tribology, 235
- Tripropellant. *See* Propellant
- True anomaly, 64
- True longitude, 67
- Tsander, Fridrikh. *See* Zander, Friedrich
- Tsiolkovsky rocket equation. *See* Rocket equation
- Tsiolkovsky, Konstantin, 2, 178, 397
- Tundra orbit, 91
- Turbulent, 105
- laminar-turbulent transition, 105
- Two-body problem, 62
- equation of motion, 63
 - solution of, 62
- Two-line element, 30
- BSTAR, 30
- U**
- Ultraviolet (UV), 39
- Uncertainty, 708
- quantification, 708
- Unified Modeling Language (UML). *See* Programming language
- Universal gravitation constant. *See* Gravitational constant
- UoSAT (spacecraft), 335, 348
- US Department of Defense Gravity Experiment (DODGE) satellite, 335
- User segment. *See* Payload data ground segment
- USML. *See* International Traffic in Arms Regulations (ITAR)
- V**
- V&V. *See* Verification & validation
- V-2 (rocket), 4
- van Allen belts, 42
- Variable Specific Impulse Magnetic Rocket. *See* VASIMR
- VASIMR (Variable Specific Impulse Magnetic Rocket), 303
- Vega (rocket), 202, 283, 287
- Vega (spacecraft), 99, 532
- Vehicle suspension
- Rocker-Bogie, 547
- Venera (spacecraft), 99
- Verification & validation, 155, 702
- acceptance, 210, 214
 - accreditation, 157
 - acoustic test, 215
 - burst pressure tests, 217
 - definition of, 156
 - deployable structures, 240
 - dynamic test loads, 208
 - flight acceptance, 210
 - flight software. *See* flight software, testing
 - ground segment, 586
 - inflatable, 245
 - in-orbit testing (IOT), 577
 - mechanisms, 237
 - micro-vibration testing, 216
 - modal survey test, 215
 - proto-flight, 210
 - qualification, 210, 214
 - random test, 215
 - safety factor, 210
 - shock test, 216
 - sine vibration test, 215
 - static load test, 214
 - structure, 211
 - system verification test (SVT), 591
 - test plan, 155
 - thermal balance testing, 393

- thermal control, 392
- thermal test facilities, 393
- thermal-vacuum testing, 393
- thermo-elastic testing, 216
- ultimate, 210
- yield, 210
- Vernal equinox, 66
- Verne, Jules, 1, 2, 7
- Vernier thrusters, 173
- Viasat-1 (spacecraft), 14
- Vibration
 - control, 218, 554
 - isolation, 218
- Vibrational excitation, 109, 113
- Vieille, Paul, 187
- Vieille's law, 285
- View factors, 378
- Viking (spacecraft), 100, 546
 - descent propulsion system, 531
- Visual odometry, 563
- Vis-viva equation, 65
- von Braun, Wernher, 3, 179
- von Kármán ellipsoid, 25, 657
- von Kármán, Theodore, 25
- Voskhod (spacecraft), 101
- Vostok (spacecraft), 8, 101
- Voyager (spacecraft), 6, 17, 55
 - flight software, 472
- Vulcain 2, 185

- W**
- Warner diagram, 176
- Waste management, 504
- Water management, 494
 - microbial contamination, 494
 - microbial monitoring, 494
 - recycling, 499
- Waterfall model. *See* Systems engineering, waterfall model
- Waverider, 103
- Weather satellite. *See* Meteorology
- Welding. *See* Joining
- Weststar (spacecraft), 398
- Wheel-terrain interactions. *See* Terramechanics
- Whiffle tree, 214
- Whipple shield, 52
- Wide-field Infrared Survey Explorer (WISE), 129
- Wien's displacement law, 377
- Wind tunnel, 109
- Wire Gage number, 259
- Work breakdown structure (WBS), 160

- X**
- X-37, 101, 515, 520
- X-38. *See* International Space Station, Crew Return Vehicle (CRV)
- Xeus (spacecraft), 27
- XMM-Newton (spacecraft), 345
- X-Prize, 300
- X-rays, 40

- Y**
- Yuhangyuan, 12

- Z**
- Zander, Friedrich, 179
- Zero point of longitude, 66
- Ziegler-Nichols method. *See* Controller, proportional-integral-derivative (PID)
- Zodiacal light, 50