

Terry D. Oswalt
Editor-in-Chief

Ian S. McLean
Volume Editor

Planets, Stars and Stellar Systems

VOLUME 1

Telescopes and Instrumentation

Planets, Stars and Stellar Systems

Telescopes and Instrumentation

Terry D. Oswalt (Editor-in-Chief)

Ian S. McLean (Volume Editor)

Planets, Stars and Stellar Systems

Volume 1: Telescopes and Instrumentation

With 251 Figures and 45 Tables

Editor-in-Chief

Terry D. Oswalt
Department of Physics & Space Sciences
Florida Institute of Technology
University Boulevard
Melbourne, FL, USA

Volume Editor

Ian S. McLean
Department of Physics & Astronomy
University of California, Los Angeles
Los Angeles, CA, USA

ISBN 978-94-007-5620-5 ISBN 978-94-007-5621-2 (eBook)
ISBN 978-94-007-5622-9 (print and electronic bundle)
DOI 10.1007/978-94-007-5621-2

This title is part of a set with
Set ISBN 978-90-481-8817-8
Set ISBN 978-90-481-8818-5 (eBook)
Set ISBN 978-90-481-8852-9 (print and electronic bundle)

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012953926

© Springer Science+Business Media Dordrecht 2013

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)

Series Preface

It is my great pleasure to introduce “Planets, Stars, and Stellar Systems” (PSSS). As a “Springer Reference”, PSSS is intended for graduate students to professionals in astronomy, astrophysics and planetary science, but it will also be useful to scientists in other fields whose research interests overlap with astronomy. Our aim is to capture the spirit of 21st century astronomy – an empirical physical science whose almost explosive progress is enabled by new instrumentation, observational discoveries, guided by theory and simulation.

Each volume, edited by internationally recognized expert(s), introduces the reader to a well-defined area within astronomy and can be used as a text or recommended reading for an advanced undergraduate or postgraduate course. Volume 1, edited by Ian McLean, is an essential primer on the tools of an astronomer, i.e., the telescopes, instrumentation and detectors used to query the entire electromagnetic spectrum. Volume 2, edited by Howard Bond, is a compendium of the techniques and analysis methods that enable the interpretation of data collected with these tools. Volume 3, co-edited by Linda French and Paul Kalas, provides a crash course in the rapidly converging fields of stellar, solar system and extrasolar planetary science. Volume 4, edited by Martin Barstow, is one of the most complete references on stellar structure and evolution available today. Volume 5, edited by Gerard Gilmore, bridges the gap between our understanding of stellar systems and populations seen in great detail within the Galaxy and those seen in distant galaxies. Volume 6, edited by Bill Keel, nicely captures our current understanding of the origin and evolution of local galaxies to the large scale structure of the universe.

The chapters have been written by practicing professionals within the appropriate sub-disciplines. Available in both traditional paper and electronic form, they include extensive bibliographic and hyperlink references to the current literature that will help readers to acquire a solid historical and technical foundation in that area. Each can also serve as a valuable reference for a course or refresher for practicing professional astronomers. Those familiar with the “Stars and Stellar Systems” series from several decades ago will recognize some of the inspiration for the approach we have taken.

Very many people have contributed to this project. I would like to thank Harry Blom and Sonja Guerts (Sonja Japenga at the time) of Springer, who originally encouraged me to pursue this project several years ago. Special thanks to our outstanding Springer editors Ramon Khanna (Astronomy) and Lydia Mueller (Major Reference Works) and their hard-working editorial team Jennifer Carlson, Elizabeth Ferrell, Jutta Jaeger-Hamers, Julia Koerting, and Tamara Schineller. Their continuous enthusiasm, friendly prodding and unwavering support made this series possible. Needless to say (but I’m saying it anyway), it was not an easy task shepherding a project this big through to completion!

Most of all, it has been a privilege to work with each of the volume Editors listed above and over 100 contributing authors on this project. I’ve learned a lot of astronomy from them, and I hope you will, too!



January 2013

Terry D. Oswalt
General Editor

Preface to Volume 1

From the first appearance of the telescope in 1608, better telescopes have led to more astronomical discoveries. Every improvement in technology helps to provide answers to old questions. Inevitably, the new observational methods uncover a host of new questions, which in turn drives the quest for even better equipment. We strive for deeper surveys, fainter limits, better resolution, larger statistical samples, and broader spectral response. Of course, the ground-based telescopes of today are enormous compared to those of Galileo. Computer control of optical surfaces has changed the way telescopes are built. By finding ways to compensate for atmospheric turbulence, very large telescopes can operate at their diffraction limit. And today, powerful space telescopes span the entire electromagnetic spectrum.

In the late nineteenth century, when dry, gelatin-based photographic emulsions became routinely available, astronomers lost no time in putting them to use. The photographic process was more accurate and more sensitive than the human eye. With the invention of the charge-coupled device in 1969, the photographic plate gave way to digital imaging devices. Today, all modern astronomical research is carried out with photoelectronic equipment that converts radiant energy into electrical signals which can be digitized for immediate storage and manipulation in a computer.

A modern observatory, on the ground or in space, and irrespective of its wavelength domain, requires an enormous breadth of engineering, scientific, and managerial skills to operate efficiently and produce the very best results. This volume provides an introduction to telescopes and instrumentation for all kinds of astronomy. Each chapter has been written by pioneers and practicing professionals.

The volume begins with a general introduction to telescopes written by David Silva and myself in order to provide a basis for the articles that follow. Survey and robotic telescopes are covered in ● Chap. 2 by Przemysław Woźniak. Next, each of the three major groundbreaking technologies for very large telescopes is described by the people who played leading roles in their development. Segmented mirror telescopes are covered by Jerry Nelson and Gary Channan; honeycomb mirrors for large telescopes are explained by John Hill, Buddy Martin, and Roger Angel; actively supported thin-mirror telescopes are covered by Lothar Noethe and Ray Wilson. Theo ten Brummelaar and Hal McAlister provide an excellent review of optical and infrared interferometry. Ron Ekers and Tom Wilson describe radio telescopes while Jonas Zmuidzinas, Tom Philipps, and Steve Padin focus on sub-millimeter and millimeter telescopes. Moving to space, there is an overview of UV/optical/IR telescopes by Matt Mountain, Erin Elliott, and Marc Postman. Very high-energy telescopes are reviewed by Elizabeth Hays, and telescopes for the detection of the cosmic microwave background are covered by Lyman Page, Michael Niemack, and Shaul Hanany. A general introduction to astronomical instruments and detectors is given in a chapter written by myself with colleagues James Larkin and Michael Fitzgerald. A more detailed chapter on silicon imaging devices is provided by Paul Jorden, and infrared detectors are described in greater depth by Erick Young, Rebecca Bernstein and Steve Shtetman

close this version of the volume with a review of astronomical spectrographs. Other topics are under development and will be added once the volume goes on-line.

It has been a privilege to work with each of the contributors to this volume, as well as our General Editor Terry Oswalt and the entire staff of Springer.

Ian S. McLean
Los Angeles, CA
USA

Editor-in-Chief



Dr. Terry D. Oswalt
Department Physics & Space Sciences
Florida Institute of Technology
150 W. University Boulevard
Melbourne, Florida 32901
USA
E-mail: toswalt@fit.edu

Dr. Oswalt has been a member of the Florida Tech faculty since 1982 and was the first professional astronomer in the Department of Physics and Space Sciences. He serves on a number of professional society and advisory committees each year. From 1998 to 2000, Dr. Oswalt served as Program Director for Stellar Astronomy and Astrophysics at the National Science Foundation. After returning to Florida Tech in 2000, he served as Associate Dean for Research for the College of Science (2000–2005) and interim Vice Provost for Research (2005–2006). He is now Head of the Department of Physics & Space Sciences. Dr. Oswalt has written over 200 scientific articles and has edited three astronomy books, in addition to serving as Editor-in-Chief for the six-volume Planets, Stars, and Stellar Systems series.

Dr. Oswalt is the founding chairman of the Southeast Association for Research in Astronomy (SARA), a consortium of ten southeastern universities that operates automated 1-meter class telescopes at Kitt Peak National Observatory in Arizona and Cerro Tololo Interamerican Observatory in Chile (see the website www.saraobservatory.org for details). These facilities, which are remotely accessible on the Internet, are used for a variety of research projects by faculty and students. They also support the SARA Research Experiences for Undergraduates (REU) program, which brings students from all over the U.S. each summer to participate one-on-one with SARA faculty mentors in astronomical research projects. In addition, Dr. Oswalt secured funding for the 0.8-meter Ortega telescope on the Florida Tech campus. It is the largest research telescope in the State of Florida.


Dr. Oswalt's primary research focuses on spectroscopic and photometric investigations of very wide binaries that contain known or suspected white dwarf stars. These pairs of stars, whose separations are so large that orbital motion is undetectable, provide a unique opportunity to explore the low luminosity ends of both the white dwarf cooling track and the main sequence; to test competing models of white dwarf spectral evolution; to determine the space motions, masses, and luminosities for the largest single sample of white dwarfs known; and to set a lower limit to the age and dark matter content of the Galactic disk.

Volume Editor



Dr. Ian S. McLean

Department of Physics & Astronomy
University of California
475 Portola Plaza
Los Angeles, CA 90095-1547
USA
E-mail: mclean@astro.ucla.edu

Dr. McLean has been a member of the faculty at UCLA and director of the UCLA Infrared Laboratory for Astrophysics since 1989. He has served as vice chair for astronomy since 2009. The IR Lab at UCLA is well known for the development of many astronomical instruments for the Keck Observatory and other telescopes. Dr. McLean is one of the world's leading authorities on the application of electronic imaging systems to advanced astronomical instrumentation. He has written over 300 articles and his current book,  *Electronic Imaging in Astronomy: Detectors and Instrumentation*, published by Springer is in its second edition.

Dr. McLean received his B.Sc. (Hons) in physics and astronomy and his Ph.D. in astronomy from Glasgow University (UK) in 1971 and 1974, respectively. While a member of staff at the Royal Observatory Edinburgh from 1979–1989, he developed the first CCD-based imaging spectro-polarimeter and the first facility-class camera for the 3.8-m UK Infrared Telescope to use infrared arrays. Since joining UCLA, his lab has delivered or assisted in all of the currently operational infrared instruments at the W. M. Keck Observatory. Dr. McLean was the principal investigator for the NIRSPEC and MOSFIRE instruments at Keck, the twin-channel infrared camera at Lick Observatory, and the FLITECAM instrument for NASA's Stratospheric Observatory for Infrared Astronomy. He served on the Science Steering Committee for the W. M. Keck Observatory for 10 years and he is an associate director of the University of California Observatories. Dr. McLean is a former president of IAU Commission 25 (Photometry and Polarimetry) and a former president of IAU Commission 9 (Instrumentation and Techniques). He is a member of the American Astronomical Society (AAS) and the Society of Professional Instrument Engineers (SPIE). His research interests are broad. From his early career, he is known for many discoveries involving the intrinsic polarization of starlight. Currently, his main research involves the study of substellar mass objects (brown dwarfs), star-forming regions, the galactic center, and star formation in high-redshift galaxies. His web site bdssarchive.org provides a public database of infrared spectra for low-mass stars and brown dwarfs.

Table of Contents

Series Preface	v
Preface to Volume 1	vii
Editor-in-Chief	ix
Volume Editor	xi
List of Contributors	xv

Volume 1

1 Introduction to Telescopes	1
<i>David Silva · Ian S. McLean</i>	
2 Robotic and Survey Telescopes	43
<i>Przemysław Woźniak</i>	
3 Segmented Mirror Telescopes	99
<i>Jerry Nelson · Terry Mast · Gary Chanan</i>	
4 Honeycomb Mirrors for Large Telescopes	137
<i>John Hill · Hubert Martin · Roger Angel</i>	
5 Active Thin-Mirror Telescopes	185
<i>Lothar Noethe · Ray Wilson</i>	
6 Optical and Infrared Interferometers	241
<i>Theo A. ten Brummelaar · Harold A. McAlister</i>	
7 Submillimeter Telescopes	283
<i>Thomas G. Phillips · Stephen Padin · Jonas Zmuidzinas</i>	
8 Radio Telescopes	315
<i>Ron Ekers · Thomas L. Wilson</i>	
9 Space Telescopes in the Ultraviolet, Optical, and Infrared (UV/O/IR)	361
<i>Erin Elliott · Matt Mountain · Marc Postman · Anton Koekemoer · Leonardo Ubeda · Mario Livio</i>	

10 CMB Telescopes and Optical Systems	431
<i>Shaul Hanany · Michael D. Niemack · Lyman Page</i>	
11 Very-High-Energy Gamma-Ray Telescopes	481
<i>Elizabeth Hays</i>	
12 Instrumentation and Detectors	507
<i>Ian S. McLean · James Larkin · Michael Fitzgerald</i>	
13 Silicon-Based Image Sensors	541
<i>Paul R. Jorden</i>	
14 Long-Wavelength Infrared Detectors	565
<i>Erick T. Young</i>	
15 Astronomical Spectrographs	587
<i>Rebecca A. Bernstein · Stephen A. Sheckman</i>	
Index	619

List of Contributors

Roger Angel

Department of Astronomy
University of Arizona
Steward Observatory
Tucson, AZ
USA

Rebecca A. Bernstein

Astronomy and Astrophysics
Department/UC Observatories
UC Santa Cruz
Santa Cruz, CA
USA

Theo A. ten Brummelaar

Mount Wilson Observatory
The CHARA Array of Georgia State
University
Mount Wilson, CA
USA

Gary Chanan

University of California
Irvine, CA
USA

Erin Elliott

Space Telescope Science Institute
Baltimore, MD
USA

Ron Ekers

Australia Telescope National Facility
CSIRO Astronomy and Space Science
Epping, NSW
Australia

Michael Fitzgerald

Department of Physics and Astronomy
University of California
Los Angeles, CA
USA

Shaul Hanany

School of Physics and Astronomy
University of Minnesota
Minneapolis, MN
USA

Elizabeth Hays

NASA Goddard Space Flight Center
Greenbelt, MD
USA

John Hill

Large Binocular Telescope Observatory
University of Arizona
Steward Observatory
Tucson, AZ
USA

Paul R. Jorden

e2v technologies plc
Chelmsford
Essex
UK

Anton Koekemoer

Space Telescope Science Institute
Baltimore, MD
USA

James Larkin

Department of Physics and Astronomy
University of California
Los Angeles, CA
USA

Mario Livio

Space Telescope Science Institute
Baltimore, MD
USA

Hubert Martin

Department of Astronomy
University of Arizona
Steward Observatory
Tucson, AZ
USA

Terry Mast

UC Observatories/Lick Observatory
Santa Cruz, CA
USA

Harold A. McAlister

Department of Physics and Astronomy
Georgia State University
Atlanta, GA
USA

Ian S. McLean

Department of Physics and Astronomy
University of California
Los Angeles, CA
USA

Matt Mountain

Space Telescope Science Institute
Baltimore, MD
USA

Jerry Nelson

UC Observatories/Lick Observatory
Santa Cruz, CA
USA

Michael D. Niemack

National Institute of Standards and
Technology
University of Colorado
Boulder, CO
USA

Physics Department
Cornell University
Ithaca, NY
USA

Lothar Noethe

European Southern Observatory ESO
Garching
Germany

Stephen Padin

California Institute of Technology
Pasadena, CA
USA

Lyman Page

Department of Physics
Princeton University
Princeton, NJ
USA

Thomas G. Phillips

Physics, Mathematics & Astronomy
California Institute of Technology
Pasadena, CA
USA

Marc Postman

Space Telescope Science Institute
Baltimore, MD
USA

Stephen A. Sackett

Observatories of the Carnegie Institution
Pasadena, CA
USA

David Silva

National Optical Astronomy Observatory
(NOAO)
Tucson, AZ
USA

Leonardo Ubeda

Space Telescope Science Institute,
Baltimore, MD
USA

Ray Wilson

European Southern Observatory ESO
Garching
Germany

Thomas L. Wilson

Naval Research Laboratory
Washington, DC
USA

Przemysław Woźniak

Los Alamos National Laboratory
Los Alamos, NM
USA

Erick T. Young

Director, SOFIA Science Mission Operations
NASA Ames Research Center
Moffett Field, CA
USA

Jonas Zmuidzinas

George W. Downs Laboratory of Physics
California Institute of Technology
Pasadena, CA
USA

1 Introduction to Telescopes

David Silva¹ · Ian S. McLean²

¹National Optical Astronomy Observatory (NOAO), Tucson, AZ, USA

²Department of Physics and Astronomy, University of California, Los Angeles, CA, USA

1	Preamble	2
1.1	What Is a Telescope?.....	2
1.1.1	General Description.....	2
1.1.2	EMR Wavelength as Key Design Driver.....	3
1.1.3	Telescopes in the Observatory Context.....	6
1.2	Geometric Optics and Idealized Telescopes.....	7
1.2.1	Basic Principles.....	7
1.2.2	Refracting Telescopes.....	12
1.2.3	Reflecting Telescopes.....	14
1.2.4	Grazing Incident Telescopes.....	17
1.3	Physical Optics and Image Formation.....	18
1.3.1	Ideal Images.....	18
1.3.2	Image Errors from Optical Elements.....	21
1.3.3	Image Errors from Atmospheric Turbulence.....	23
1.3.4	Parameterization of Image Quality and Wavefront Error.....	25
1.3.5	Aperture Synthesis.....	27
1.4	Ancillary Systems.....	29
1.4.1	Telescope Mounts and Motion Control.....	29
1.4.2	Active Optics and Optics Control.....	33
1.4.3	Adaptive Optics and Turbulence Compensation.....	35
1.4.4	Enclosures and Protection from Inclement Environmental Conditions.....	38
1.5	Looking Forward.....	40
	References	40
	Selected References (Books).....	40
	Selected Technical/Semi-technical References (Journal Articles).....	41
	Selected Technical or Semi-technical Publications (Web-Based).....	42
	SPIE Conference Proceedings.....	42

1 Preamble

Since the dawn of consciousness, humanity has sought to make sense of the surrounding universe by observing the world and recording what is seen. Paleolithic figurines and cave paintings testify silently to early attempts to find meaning in what seemed to be chaos. While some phenomena could be explored by multiple senses, the sky above with its immovable and movable objects was only observable by vision.

Today, our ability to observe the universe visually has been extended enormously by the telescope, a device to collect and focus *electromagnetic radiation (EMR)* on to a recording device. While the original recording device was a combination of the human eye to see and the human hand to draw, electronic imaging and recording now dominate. The handheld telescopes of Galileo have been replaced by sophisticated machines on the ground and in space that can be used to capture radiation across the entire electromagnetic spectrum to study objects and events distant in space and time.

This chapter starts with a general description of the astronomical telescope, what drives its design, and how it fits within an observatory context. Principles of geometric optics are used to introduce ideal telescopes while principles of physical optics are used to describe how images are formed and deformed. Important ancillary systems are also called out and described. The overall intent is to provide enough introductory background and nomenclature to allow readers to efficiently seek more details in the literature and to provide a foundation for understanding the more detailed reviews found elsewhere in this volume.

1.1 What Is a Telescope?

1.1.1 General Description

A telescope is a device for observing distant objects by collecting their emitted EMR and concentrating it onto a recording device. Most but not all telescopes create enlarged (magnified) images of those distant objects. Image brightness depends on how much EMR power is collected. Since most astronomical objects of interest are faint, maximizing telescope collecting area and throughput without degrading image quality (or exhausting financial resources!) is always desired. In the nomenclature of signal or information processing, telescopes are devices for collecting and processing EMR.

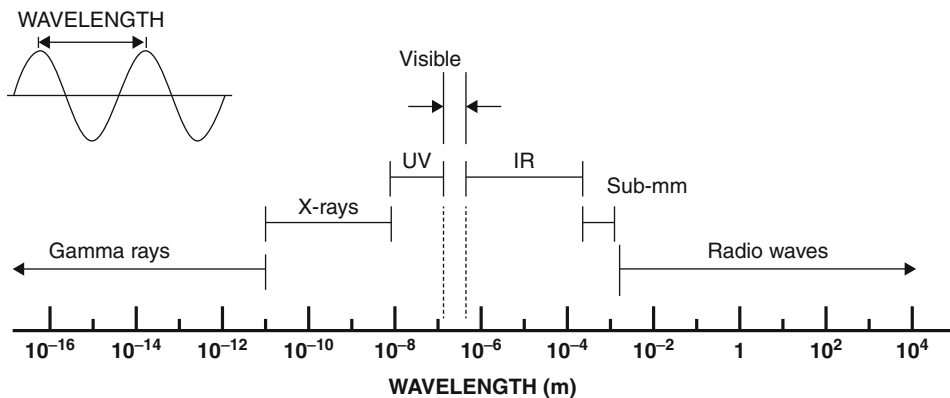
Mechanically, a telescope consists of one or more optical elements, each mounted within an individual mechanical structure designed to keep the optical elements safe and properly shaped. In turn, those structures are mechanically connected within a structure designed to maintain their alignment relative to a common axis. The first optical element to collect light is called the primary, the second is called the secondary, and so on. The primary elements in the first astronomical telescopes were refractive (lenses) but most professional telescopes now use reflective optics (mirrors). Secondary optics and so on can be a combination of reflective and refractive optics. Originally passive, most modern professional optical systems have active mechanisms to compensate for mechanical deformation caused by time-variable thermal or gravity loading effects. The surface (collecting) area of the primary determines the amount of EMR energy collected. Energy delivered to the recording device is less than collected energy due to a combination of optical beam obstructions and imperfect reflection or transmission by optical elements.

Telescopes fall into three general types: *refractor* (all lenses), *reflector* (all mirrors, although lens systems are often used as so-called field correctors), and *grazing incidence* (a special case of reflector for EMR with very small wavelength, i.e., x-ray). Because the main EMR collecting elements of such systems are spatially continuous, they are known as filled aperture telescopes. Unfilled aperture telescopes, otherwise known as spatial interferometers, combine the input of multiple telescopes to create images with spatial resolution equivalent to a filled aperture telescope with diameter equal to the separation of the individual telescopes in the interferometer. Each of these types is discussed in more detail later.

1.1.2 EMR Wavelength as Key Design Driver

The most important telescope design criterion is wavelength (frequency) of incident EMR. Why? Because the wavelength of EMR emitted by astronomical objects is intimately connected with the physical mechanisms that create it, and in turn, it is those mechanisms that astronomers want to investigate. The extremely hot (millions of degrees K) gas that fills the gravitational potential wells of galaxy clusters or surrounds the Sun emits EMR predominately at x-ray (0.01–10 nm) wavelengths. At the other end of the temperature scale, the extremely cold (2.7 K) cosmic microwave background emits EMR at microwave (1 mm–1 m) wavelengths. All celestial phenomena emit a combination of continuum and spectral line radiation. The former is often emitted by thermal mechanisms whose EMR spectrum can be closely approximated by a blackbody spectrum $I(\lambda)$ parameterized by a specific temperature. Other physical origins of continuum radiation exist and have their own characteristic $I(\lambda)$ forms (e.g., synchrotron radiation). Emission lines are created by photon emission due to electronic and nuclear energy state transitions in atoms and molecules. Conversely, absorption lines are created by photon absorption by the same energy state transitions. To understand the physical underpinnings of various astrophysical phenomena, it is often necessary to observe continuum and spectral line radiation at several widely separated wavelengths (🔗 Fig. 1-1).

Many astronomers want to study objects in the distant universe. Because the speed of light is finite and the universe is expanding, the wavelength of EMR emitted by the object is stretched

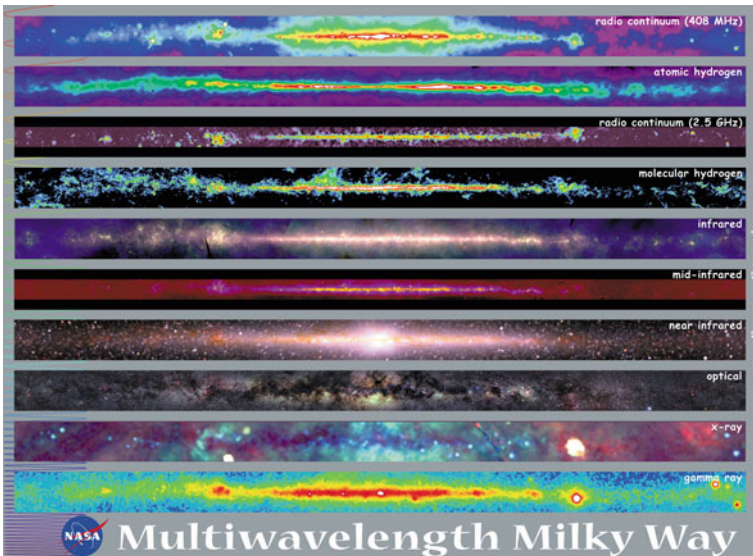


■ Fig. 1-1
Electromagnetic spectrum

during its journey to Earth and is said to become redder. Therefore, spectral features critical for probing physical conditions and mechanisms at their creation point must be observed at longer wavelengths. A profound example of this effect is the cosmic microwave background, created about 380,000 years after the Big Bang and emitted at $\lambda = 1.69$ nm but observed at $\lambda = 1.84$ mm (redshift $z = 1,089$). To study the geometry of the universe itself, redshift measurements of large samples of distant objects can be collected to probe its expansion history as well as how mass (matter) is distributed.

Before proceeding, recall that EMR has wavelike properties where wavelength (distance between waves) and frequency (waves per second) are related by the speed of light for a given medium ($c = \lambda\nu$) as well as particle-like properties where quantized energy per photon is equal to Planck's constant multiplied by frequency ($E = h\nu$). In a very deep sense, EMR detection *energy* is a function of its particle-like properties while EMR detection *location* is a function of its wave-like properties. For historical reasons related to detection and recording technology, astronomers emphasize different wavelength, frequency, and energy units in different wavelength bands. Gamma ray and x-ray astronomers tend to use energy units (e.g., keV), optical-infrared astronomers use wavelength units (e.g., Ångstroms, nanometers, microns), and radio astronomers tend to use frequency units (e.g., MHz, GHz). In this chapter, wavelength units are used throughout for consistency (🔗 Fig. 1-2).

Selected based on science goals, chosen wavelength drives two critical decisions: optical design (especially surface quality of all optical elements) and location (especially altitude above sea level). Important related decisions include desired image quality and field of view. Such decisions must be moderated by fiscal reality – just because something is technologically possible (or plausible) does not mean it is affordable.



🔗 Fig. 1-2

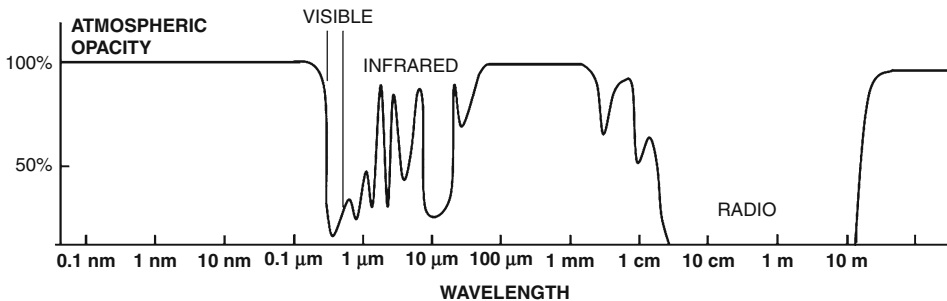
Milky Way galaxy observed at multiple wavelengths (Image credit: NASA/Goddard Space Flight Center)

Following these decisions, all telescopes are designed around a primary optical element that must collect EMR of a given wavelength. To concentrate EMR usefully, the primary optical element must be significantly larger and smoother than the wavelength of the incident EMR. For filled aperture systems, primary element size also determines achievable spatial resolution and amount of EMR power collected, both important given that most celestial objects of interest are faint and small in angular size.

The terrestrial atmosphere attenuates or completely blocks broad ranges of the EMR spectrum. While this is fortunate for humans in general, it creates a significant challenge for astronomers. At wavelengths less than $0.3\ \mu\text{m}$, oxygen (O_2) and ozone (O_3) are the dominant absorbers, and no EMR reaches the ground. Research at these wavelengths requires telescopes flown at stratospheric altitudes or completely outside Earth's atmosphere, adding significant complexity and expense (● Fig. 1-3).

Between 0.3 and $2\ \mu\text{m}$, wide atmospheric transmission windows exist, moderated by absorption from dust (e.g., volcanic), ozone (below $0.32\ \mu\text{m}$), oxygen (most notably in the so-called A and B bands at 0.760 and $0.688\ \mu\text{m}$, respectively), various molecular species (e.g., hydroxide, OH^-), and, especially, water vapor. Telescopes designed to work at these wavelengths are typically located on remote, high ($2,000\text{-m}$ or greater) mountain tops to reduce interference from artificial EMR and to rise above the inversion layer that increases local atmospheric turbulence, low altitude, and dust interference. At these wavelengths, astronomers also seek dry, cloud-free sites. High-altitude dust from, for example, volcanoes is an additional, unpredictable source of attenuation for celestial EMR. At these wavelengths, atmospheric dust is large relative to observed λ and to first order absorbs uniformly with wavelength. Working above the atmosphere removes atmospheric turbulence, leading to best possible image quality but also highest expense.

Between 2 and $2,000\ \mu\text{m}$, transmission windows are defined primarily by the molecular absorption properties of water vapor and carbon dioxide (CO_2) with lesser absorption contributions coming from oxygen, ozone, methane (CH_4), and nitrous oxide (N_2O). Since the fractional water vapor content falls rapidly with altitude, researchers working at these wavelengths seek very high altitude sites ($4,000\ \text{m}$ and higher), use balloons or airplanes above the troposphere ($\sim 20,000\ \text{m}$ and higher), or go outside of the atmosphere completely. Higher altitude means lower temperatures, which is also important to reduce the thermal background that can overwhelm the celestial signal being investigated.



■ Fig. 1-3
EMR transmission of terrestrial atmosphere vs. wavelength

None of the windows between 0.29 and 2,000 μm are completely transparent. The amount of absorption is a function of current atmospheric dust content (a function of altitude above sea level), absorber scale height (again, a function of altitude above sea level), and path length through atmosphere (increasing as the telescope points further away from zenith). Path length through the atmosphere can be parameterized as *airmass* (X), the secant of the angle between the observed object and the apparent zenith of the observer. Airmass is unity when the telescope is pointed at the zenith and increases as the telescopes points toward the horizon.

Between 2,000 μm (0.02 m) and 20 m, the atmosphere is essentially transparent. Free electron scattering in the ionosphere blocks all celestial EMR for wavelengths >20 m.

1.1.3 Telescopes in the Observatory Context

The stereotypical image of an astronomical observatory is a large enclosure with a hemispheric dome located on top of a remote mountain. The truth behind this stereotype is that all telescopes are surrounded by various ancillary systems that together allow the telescope to function properly. In the ensemble, the telescope and these systems are known as an *observatory*. While short descriptions of typical ancillary systems are provided here, more details are given later.

The *telescope mount* is used to point the telescope toward an astronomical object and then move the telescope to compensate for the rotation of Earth, motion around Earth, or motion through space. These actions are commonly known as target acquisition and tracking. The mount often provides various services (e.g., communication, electrical power, coolant) needed for telescope and instrument operation. For ground-based telescopes, two kinds of two-axis mounts are common: equatorial and altitude-azimuth (alt-az). For equatorial mounts, one axis of motion is aligned with Earth's axis of rotation while the other axis is orthogonal to that (i.e., parallel to Earth's equator projected on the celestial sphere). For alt-az telescopes, one axis of motion is perpendicular to the surface of Earth and the other is orthogonal to that (i.e., parallel to Earth's surface).

The *telescope control system* (TCS) accepts positional coordinates of an astronomical object and then issues appropriate motion commands to the telescope mount. Telescope control systems are embedded within more extensive observatory control systems that coordinate the actions all telescope and observatory systems. The TCS is often associated with a telemetry system that collects and stores data about system status.

EMR collected by telescopes are projected into *instruments* capable of recording received energy (power) or created images. Instrument nomenclature follows function and wavelength band, using historical conventions. For example, instruments that record energy are usually called receivers for radio telescopes but bolometers for infrared telescopes. Spectrographs (spectrometers) are instruments that disperse received EMR. Instruments can be physically connected to the telescope or located nearby.

A ground-based telescope, its mount, and its instruments must be protected from weather to function properly. As the ensemble is too large to be moved under a shelter when not in use, it is surrounded by an *enclosure*. Modern enclosure design strives to balance protection from the elements (e.g., telescope motion induced by wind buffeting, optical surface damage from rain or dust) and the need to minimize thermal turbulence within the enclosure that can degrade delivered image quality. Exact details depend on the telescope wavelength range. For example, as λ increases into the submillimeter range and beyond, problems related to internal turbulence and optical surface damage become less important. In the radio band, telescopes are left exposed

to the elements all the time. Space-based telescopes do not have enclosures per se but sometimes have sunshades to minimize heat loading from solar illumination and to prevent damage to instrumentation from inadvertent observation of the bright sources such as the Sun.

Ground-based observatories have *support facilities* for activities such as performance monitoring, correction and preventive maintenance, utility delivery (e.g., power, water, communications), implementation of new capabilities, and data processing and archiving. Space-based facilities have a similar set of support facilities. Of course, in situ spacecraft maintenance and improvement is not possible for most such observatories, with the obvious exception of the Hubble Space Telescope.

Finally, the *observatory staff* is a group of people that operate, maintain, and extend observatory capabilities. A broad mix of scientific, technical, and administrative skill is needed. In a mature professional observatory, the annual cost of operation is usually dominated by labor (staff) costs. Over the lifetime of an observatory, operations costs eventually dominate the original construction costs. Therefore, minimization of operations costs is an important design requirement for modern observatories.

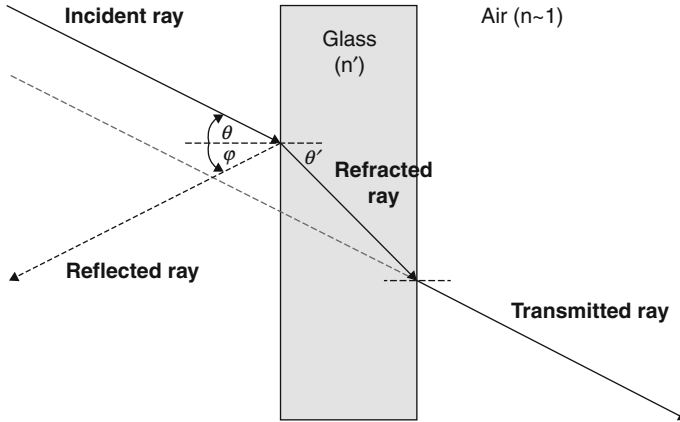
1.2 Geometric Optics and Idealized Telescopes

1.2.1 Basic Principles

Modern development of EMR wave-particle duality began with work by Huygens (1690) (wave) and Newton (1704) (particle). The wave-based unification of electricity and magnetism (Maxwell 1861), the discrete oscillator explanation of blackbody emission (Planck 1901), and the photon-based explanation of the photoelectric effect (Einstein 1905) lead ultimately to the development of quantum mechanics (Heisenberg 1925). Astronomers tend to think in terms of photons (or equivalently energy) collected per unit time per unit area per wavelength (or frequency) since that has the most direct physical connection to the phenomena being studied. Photons (often in terms of intensity or power) are also what astronomical detectors measure.

However, it is the wave-like properties of EMR that determines detection location. Hence, optical designers follow the approach first proposed by Huygens and work in terms of spatially continuous wave fronts composed of the superposition of the individual wave fronts emitted by idealized point sources. In the realm of geometric optics, such composite wave fronts are conceptualized as bundles of *rays*, arbitrarily narrow beams of light of known wavelength λ . Rays represent the direction of flow of the energy in an electromagnetic wave and are perpendicular to the wavefront. This simplification is highly appropriate when λ is small relative to the size of the intervening optical elements. However, it is cumbersome for modeling two-dimensional image creation via diffraction (interference) and image modification via aberration. Imaging is more easily conceptualized by physical optics, discussed later in this chapter (► [Fig. 1-4](#)).

Rays travel in straight lines in homogeneous media but curve in media with variable refractive indices. Rays are split when they strike the boundary between two transparent materials. One component is bent (refracted) as it crosses the boundary while the other is reflected away from the boundary. At the point of contact, the ray makes an angle of incidence θ with a line that is perpendicular or “normal” to the interface at that point. As $\theta \rightarrow 0$, incident rays are said to have normal (or nearly normal) incidence angles.



■ Fig. 1-4
Basic principles of reflection and refraction

The *law of reflection* has been known since antiquity (putatively first by Hero of Alexandria) and states that the reflected ray lies in the same plane as the incident ray and the angle of reflection (φ) relative to the normal is equal to the angle of incidence. Therefore,

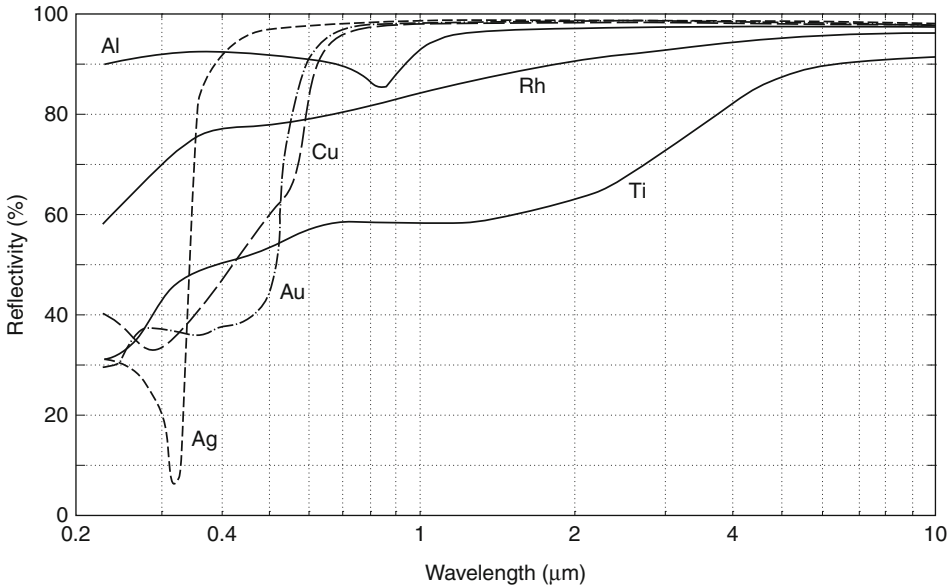
$$\theta = \varphi \quad (1.1)$$

Specular (mirror-like) reflection is produced by surfaces where all incident rays in a bundle are reflected in the same direction. As a general rule, a surface is considered to be specular (smooth) when $8h \times \cos(\theta) < \lambda$, where h is the characteristic roughness height of the material. This is called the Rayleigh criterion, not to be confused with the Rayleigh criterion for resolvability of point sources. Today, optical smoothness is usually specified by a structure function that defines h (or σ_h) as a function of measurement point separation. As surface roughness increases relative to λ , the ray bundle broadens and the reflection is said to become more diffuse.

At optical-infrared wavelengths, once a smooth surface is produced (usually on glass substrate), a thin (~ 100 nm) metal *coating* is applied to maximize reflectivity, the ratio of reflected energy to incident energy. Reflectivity as a function of wavelength varies from metal to metal (see [Fig. 1-5](#)). While aluminum is used at many professional telescopes, observatories that concentrate on observations beyond $1 \mu\text{m}$ have used silver and gold. Recently, complex multilayer coatings have been developed to better and more uniform reflectivity curves than are possible with single-layer coatings.

Energy that is not reflected is transmitted or absorbed. Absorbed energy is later reemitted as longer wavelength (lower energy) EMR. Surface material (or differences in material at interfaces) can cause the phase and/or polarization of the reflected light to be different than the incident light, as seen, for example, in visible light ($\lambda \sim 0.5 \mu\text{m}$) reflected from the surface of calm water. At small enough wavelength (e.g., x-ray, $\lambda \sim 0.001 \mu\text{m}$), conventional reflectivity at near normal incidence is negligible.

The behavior of the transmitted ray is described by the *law of refraction*, first accurately described by Ibn Sahl (891), but commonly known as Snell's Law (Snellius 1621), which states (in its common sine form after Descartes 1637) that the transmitted ray lies in the same plane as the incident ray, but moves in a different (deviated) direction such that the sine of the angle of



■ Fig. 1-5
Reflectivity for several metals as a function of wavelength

refraction (θ') divided by the sine of the angle of incidence (relative to the normal at that point) is a constant, equal to the ratio of the refractive indices (n/n') in the two materials. This result is commonly written as

$$n \sin \theta = n' \sin \theta' \quad (1.2)$$

In physical terms, n is the ratio of EMR speed v within a medium relative to the speed of light in vacuum, that is, $n = c/v$. At visible wavelengths, $n = 1-2$ for transparent materials. The change of ray direction can be understood by realizing that the first part of the wavefront to intercept the surface is slowed down relative to an adjacent front that travels further in the same time, and so the entire wavefront seems to progressively change direction. The distance between crests inside the material is now λ/n . As shown by Newton, refraction separates white light into its component colors, which implies that the refractive index varies with wavelength. The angular divergence of EMR of different wavelengths is called *dispersion*. The inverse dispersive power V of a given material is expressed as the difference in the refractive indices at two extreme wavelengths (n_1, n_2) relative to the difference between their average $\langle n \rangle$ and vacuum, that is,

$$V = \frac{n_2 - n_1}{\langle n \rangle - 1} \quad (1.3)$$

When EMR travels a distance d in a medium of refractive index n , the product nd is the *optical path* (or *path of least time*). Given the definition of n , the optical path is the distance that the EMR would travel in vacuum in the same time that it travels the distance d in the medium. Fermat first generalized this principle in the mid-1660s. The modern version of Fermat's principle is that the path taken by an EMR ray from one point to another through any set of media is such that its optical path is equal, in the first approximation, to other paths closely adjacent

to the actual one. In the language of calculus, the optical path must be a minimum, maximum, or stationary point (point of inflection) for the actual path.

Optical rays are said to be *paraxial* when the angles of reflection or refraction are small enough to satisfy the condition $\sin \theta = \theta$ (in radians). A *paraxial optical system* is a simplification where all elements are infinitely thin, rotationally symmetric and centered on a common, central *optical axis*. Only images centered on the optical axis (*on-axis images*) and created on a flat focal plane parallel to the optical elements are considered. Paraxial optics and relationships are also known as first-order optics and solutions because only the first terms of series expansions of trigonometric expressions are needed to describe basic on-axis optical properties. As described later, higher order terms in the series expansions are needed to describe off-axis image formation and distortion effects on curved focal surfaces.

From these paraxial assumptions, it is possible to derive a number of well-known formulas for lenses and mirrors. Such derivations are given in all standard textbooks on optics; thus, only a few useful results are provided here.

The *thin lens equation* in air or vacuum is written as

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{s'} \quad (1.4)$$

where f is the *focal length* (the distance from the optical element to the point where an image is formed) and s and s' are the object and image distances respectively. The distances, s and s' are called *conjugate points*. A lens is considered thin if its thickness is small in relation to its focal length. This equation assumes that the lens is in vacuum ($n = 1$) or to a first approximation air ($n \approx 1.0003$). If not, then the expression can be generalized to include the refractive indices of the mediums. For astronomical telescopes, the object distance is effectively infinite, $1/s \rightarrow 0$, and the conjugate image distance is just the focal length of the lens ($s' = f$).

Newton's form of the thin lens equation is

$$x'x = f^2 \quad (1.5)$$

where the distances are now measured relative to the focal points ($x = s - f$ and $x' = s' - f$).

The *lateral or transverse magnification* orthogonal to the optical axis is

$$m = -\frac{s'}{s} = \frac{h'}{h} \quad (1.6)$$

Here, h and h' are the object and image heights (lengths). The *longitudinal magnification* (along the optical axis) is m^2 . With rays traveling left to right through the lens (i.e., the object is to the left of the lens), the sign convention is that s is positive if on the left of the lens and s' is positive if on the right, and the negative sign for m indicates that the image is inverted when s and s' are both positive.

The largest unobstructed diameter in an optical system defines the amount of light that can be collected and is called an *aperture stop*. For filled aperture telescopes, the aperture stop is usually set by the diameter of the primary optical element. For many systems, the *field of view* (FOV) that can be imaged is not set by the aperture stop but by other obstructions. In those cases, the diameter that limits the field of view is called the *field stop*. Since obstructions also block rays admitted at the edge of the aperture stop, EMR power received decreases toward the edge of the field stop, an effect known as *vignetting*. If there are no such obstructions, the aperture and field stops are one and the same. The *entrance pupil* is an image (often virtual) of the aperture stop as seen from the entrance of the optical system. Conversely, the *exit pupil* is an image of the aperture stop as seen through the optical system.

For a thin lens, the *focal ratio* (otherwise known as the *f-number*, *f-ratio*, or *f-stop*) is the diameter of the entrance pupil (D) in terms of focal length. It is expressed as a dimensionless number, for example, $f/16$:

$$f\text{-ratio} = \frac{f}{D} \quad (1.7)$$

The *angular field of view* (φ , in radians) for a single lens can be written as

$$\varphi = 2 \times \arctan \left(\frac{1}{2 \times f\text{-ratio}} \right) \quad (1.8)$$

As noted above, the achieved field of view is often determined by a field stop that is different than the aperture stop.

Effective light collected is sometimes defined as *aperture area*, A , where

$$A = \pi \left(\frac{f}{2 \times f\text{-ratio}} \right)^2 \quad (1.9)$$

As f -ratio decreases, magnification decreases, while field of view and light concentration per unit area increase at the focal point. Elements with small focal ratios (like $f/1$) are said to be “fast” since it takes less time to collect a certain amount of light. Smaller f -ratios also imply shorter telescopes and smaller enclosures, reducing overall cost.

The paraxial simplification assumes on-axis images formed on a flat focal plane. In reality, each optical element projects its exit pupil onto a curved focal surface. As f -ratio decreases, so must focal length and the focal plane becomes more curved. For many applications, additional corrector optics must be introduced to project a flat enough focal plane onto a flat detector (e.g., a charge coupled device, CCD).

Equations similar to the thin lens equations hold for spherical mirrors in the paraxial approximation.

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{s'} = \frac{2}{R'} \quad (1.10)$$

where $f = R/2$ and R is the mirror *radius of curvature*. By convention, convex mirrors have negative radii of curvature and concave mirrors of positive radii of curvature. So, by (1.10), convex mirrors have negative focal lengths and are said to form virtual images behind the mirror to allow geometric analysis. Conversely, concave mirrors have positive focal points and form real images in front of the mirror. For astronomical telescopes, the object distance is effectively infinite, $1/s \rightarrow 0$, and therefore, $s' = f = R/2$.

The *optical power* ($P = 1/f$) of a thin lens is a measure of how much an optical element converges or diverges incident EMR. Optical power has units of diopters (inverse meters). It is related to the physical properties of a lens by the *lensmaker's formula*:

$$P = \frac{1}{f} = (n - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{t(n - 1)}{nR_1R_2} \right] \quad (1.11)$$

where R_1 and R_2 are the radius of curvature of the front surface and the back surface, respectively, and t is the central thickness of the lens. If t is less than $1/6$ of the lens diameter, then the third term can be neglected. Thick lenses can be handled by defining two principal planes where refraction occurs or by point-to-point ray tracing using computer programs.

For thick lenses, or two thin lenses separated by a distance d , power is additive, so that overall power, or *effective focal length*, is given by

$$P = P_1 + P_2 - \left(\frac{d}{n}\right) P_1 P_2 \quad (1.12)$$

Convex (positive, converging) lenses focus paraxial rays at a focal point centered on the optical axis. Concave (negative, diverging) lenses cause paraxial rays to diverge and appear to come from a focal point on the same side of the lens as the incident light. Image formation and the resultant magnification depend on the relative positions of the object and the focal points. When the image formed by the lens cannot be made visible on a screen because no EMR actually travels to or from that location, but only appears to do so, such images are said to be “virtual.” A simple example occurs when an object is placed closer to a convex lens than its focal point. The EMR appears to come from a magnified, upright virtual image on the same side of the lens as the object.

Angular magnification is defined in terms of the slope angles of the rays rather than the object and image heights. If the incident ray makes an angle u with respect to the optical axis (slope = $\tan u$) and the refracted ray makes an angle u' with the optical axis, then the angular magnification is

$$M = \frac{\tan u'}{\tan u} = \frac{s}{s'} = \frac{h}{h'} \quad (1.13)$$

More generally, magnification is often characterized in terms of the *plate scale*, the number of seconds of arc on the sky corresponding to 1 mm at the focus of the telescope:

$$\text{Plate Scale (''/mm)} = 206,265/f \quad (1.14)$$

where the focal length f is in mm. The primary element of a telescope always has the shortest focal length and therefore the largest plate scale.

The *Lagrange invariant* H is a measure of light propagation, constant across all refractions and reflections. In the paraxial limit,

$$hnu = h'n'u' \quad (1.15)$$

where h and h' are the object and image heights, n and n' are the refractive indices in object and image space, respectively (usually $n = n' = 1$), and u and u' are the (small) angles with respect to the optical axis of the same ray in object space and in image space. Astronomers often use this in its area-solid angle ($A\Omega$) form, also known as *étendue* or throughput. Because the total flux collected from a uniformly radiating source is proportional to ($A\Omega$) it follows that the Lagrange invariant is a consequence of the conservation of energy.

For a two-element system, the *effective focal length* is

$$f_{\text{eff}} = \frac{f_1 \times f_2}{f_1 + f_2 - d} \quad (1.16)$$

where f_1 and f_2 are the focal lengths of the two optical elements and d is the separation between the two elements. Recall that focal length is negative for concave (diverging) lens and mirrors.

1.2.2 Refracting Telescopes

Refracting telescopes (or simply refractors) use a primary optical element called an objective lens to form an image that is then magnified by a second optical element called an eyepiece,

which consists of one or more lenses. Refractors were the earliest telescopes, appearing first in the Netherlands in 1608. Their invention is usually credited to Lippersky, Janssen, and Adriaan van Leeuwenhoek. Galileo Galilei is credited with the first use of telescopes to observe celestial objects in 1609. Such telescopes are said to be dioptric since the objective lens is convex. In these simple systems, the objective is both the aperture stop and the entrance pupil while the final, magnified image of the objective is the exit pupil. A closed tube connects the objective and eyepiece to prevent optical rays from bypassing the objective lens (aperture stop), entering the eyepiece, and overwhelming the rays received from the astronomical object of interest. Refractor tubes were the first example of telescope baffles.

The basic properties for the refractor are illustrated in **Fig. 1-6** where the double-ended arrow marked O represents the long-focus (weak) objective lens and the smaller (strong) eyepiece lens is labeled E. For essentially infinite objects ($s \gg f$), **(1.4)** shows that the objective forms an image at its focal point ($s' = f$), which in turn becomes the object for the eyepiece lens. If the eyepiece is moved along the axis such that the real image formed by the objective lens coincides with the focal point of the eyepiece lens (separation of the lenses is then the sum of the focal lengths), then the emergent rays are parallel and the final (virtual) image is at infinity. Notice that the telescope is afocal: parallel rays in, parallel rays out, and the image is inverted. This is a Keplerian refractor (Kepler 1611). In contrast, a Galileian refractor (Galilei 1610) uses a divergent (concave) eyepiece, and the lenses are separated by the difference (not the sum) of their focal lengths. One ray in the figure is labeled as a chief ray because it passed through the center of both the entrance and exit pupils.

Refractors have several inherent problems. Most important is *chromatic aberration*. Because refractive index changes with wavelength, a single lens forms a series of images along the optical axis, that is, focal length f is a function of wavelength. The horizontal distance along the optical axis between the different focal positions is called *longitudinal chromatic aberration*. But the lateral magnification (m) must vary as well because the image height (h') depends on s' which is controlled by f . The vertical difference in image height is called *lateral chromatic aberration*. The classical method for correcting both forms of chromatic aberration is to use two lenses of different materials (historically crown glass and flint glass) in contact to make an achromatic doublet. Colors outside the corrected range can still cause a halo of color around a point source referred to as the secondary spectrum. However, each lens absorbs a significant amount of incident EMR, which varies as a function of lens material and wavelength.

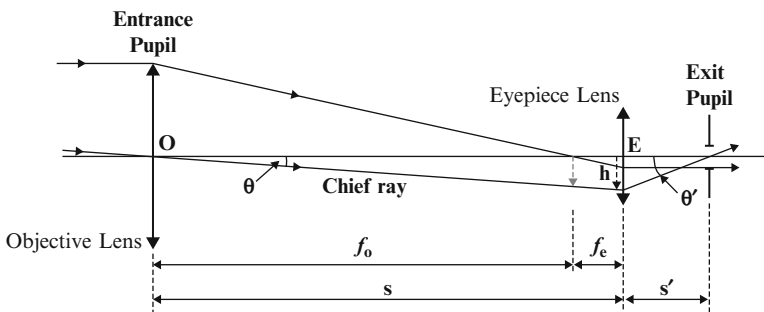


Fig. 1-6
Basic layout of refracting telescope

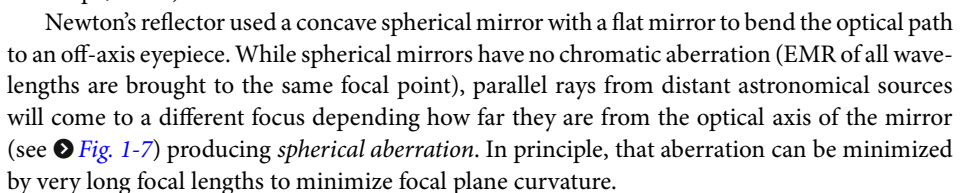
Another problem with refractors is the size and f-ratio of the objective lens. To observe fainter objects, more light must be collected, so aperture area and hence objective size must be increased. However, the manufacture of large lenses using high transmissive material without internal defects is difficult. Furthermore, since they can only be supported at the edge, large objective lenses bend (sag) due to gravity, degrading delivered image quality. Finally, because such lenses have large (slow) f-ratios, their focal lengths are long, resulting in long telescopes that also bend (sag) due to gravity. In practice, the largest refractor ever built was the Great Paris Exhibition Telescope of 1900 with an objective diameter of 1.25 m at $f/46$ – a focal length of 57 m!

1.2.3 Reflecting Telescopes

Soon after the invention of the refractor, the idea of using curved (powered) mirrors to build telescopes was discussed by Galileo, among others. Consideration of mirror-based telescopes accelerated once it was recognized that curved mirrors could be used to overcome some of the intrinsic problems of refractors. Newton is credited with building the first mirror-based telescope in 1668, although Gregory published reflector designs several years earlier (1663).

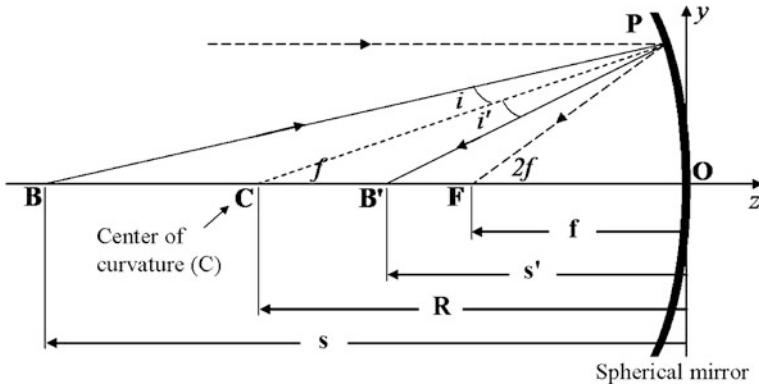
Reflecting telescopes (or simply reflectors) use one or more mirrors to form an image. Such telescopes are known as catoptric because they use curved mirrors. When refractive field curvature correctors are introduced, the combined system is said to be catadioptric. Reflectors have several key advantages over refractors, including significant reduction of chromatic aberration (since EMR of all wavelength is focused at the same point), reduced EMR scattering (due to smoother surfaces), higher end-to-end EMR transmission, ability to use additional mirrors to fold the focal length of the primary mirror into a smaller volume, and ability to build arbitrary large primary mirrors.

Today, virtually every professional astronomical telescope is a reflector of one sort or another. In comparison to the largest objective lens (1.25 m), the largest existing monolithic primary mirrors used at optical-infrared wavelengths are 8.4 m in diameter at $f/1.1$ (Large Binocular Telescope, LBT) while the largest segmented primary mirror is 10.4 m at $f/1.65$ (Gran Telescopio Canarias, GTC). Designs exist for 30 m and larger segmented optical-infrared mirrors. The largest steerable filled-aperture radio telescope has a diameter of 100 m (Robert C. Byrd Green Bank Telescope, GBT) but a nonsteerable 305-m-diameter telescope has been in operation for decades (Arecibo). A 500-m-diameter system (500 m Aperture Spherical Telescope, FAST) is under construction in China.

Newton's reflector used a concave spherical mirror with a flat mirror to bend the optical path to an off-axis eyepiece. While spherical mirrors have no chromatic aberration (EMR of all wavelengths are brought to the same focal point), parallel rays from distant astronomical sources will come to a different focus depending how far they are from the optical axis of the mirror (see  Fig. 1-7) producing *spherical aberration*. In principle, that aberration can be minimized by very long focal lengths to minimize focal plane curvature.

In practice, the problem of spherical aberration was solved using mirrors with the shapes of conic sections where any ray starting at one focus will form a perfect point image at the other (an effect called *stigmatism*). Conic sections can be described by

$$z(r) = \frac{cr^2}{1 + \sqrt{1 - (k+1)c^2r^2}} \quad (1.17)$$



■ Fig. 1-7
Reflection from a spherical surface

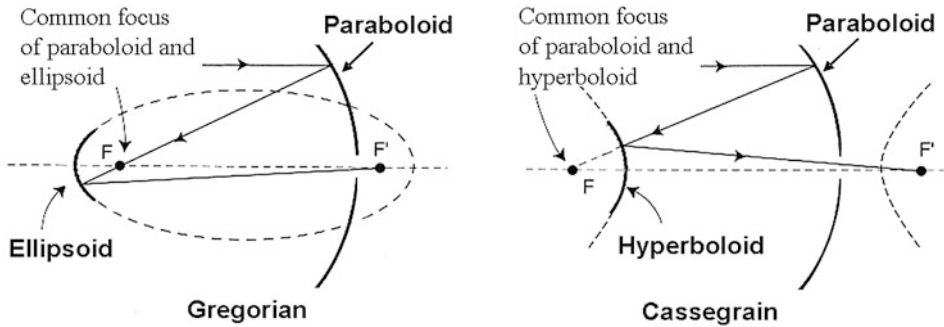
■ Table 1-1
Conic constant vs. conic surface

Conic constant	Conic surface
< -1	Hyperboloid
-1	Paraboloid
-1 to 0	Prolate ellipsoid
0	Sphere
> 0	Oblate ellipsoid

where r is the distance from the axis of rotational symmetry, (usually the optical axis) r_i is the radius of curvature of the mirror at its vertex, c is the vertex curvature of the surface perpendicular to the optical axis ($= 1/r_i$), $z(r)$ is *sag* (deflection from a flat surface perpendicular to the optical axis) as a function of r , and k is the *Schwarzschild (conic) constant* (see ▶ [Table 1-1](#)). Two-mirror *aplanatic systems* that minimize spherical and comatic aberration across large FOV can be created through proper matching of conic constants.

Gregory (1663) proposed the first such design. A *Gregorian telescope* combines a concave parabolic primary mirror with a concave ellipsoidal secondary mirror (see ▶ [Fig. 1-8](#)). Converging rays from the parabolic primary are allowed to go through the focal point of the primary (the *prime focus*) and reflect from the elliptical secondary mirror back through the hole in the primary to a (curved) focal plane.

The idea of combining a concave parabolic primary mirror with a convex hyperbolic secondary mirror appears is commonly attributed Cassegrain (see ▶ [Fig. 1-8](#)). Converging rays from the parabolic primary are allowed to go through the focal point of the primary (the *prime focus*) and reflect from the elliptical secondary mirror back through the hole in the primary to a (curved) focal plane (see ▶ [Fig. 1-8](#)). Cavalieri (1632), Mersenne (1636), and Gregory (1663) proposed similar designs. In the Cassegrain version, the hyperbolic secondary is placed to intercept optical rays before they reach the prime focus. Image quality is limited by *comatic aberration (coma)* where images become more and more comet-like as angular distance from the optical axis increases. For optical telescopes, the final focus is projected behind the primary, but radio telescopes often produce a final focus in front of the primary. In asymmetric designs,



■ Fig. 1-8

Schematic layout of Gregorian and Cassegrain reflecting telescopes

one or both mirrors may be tilted to minimize how much primary collecting area is obscured by the secondary. Magnification is simply the ratio of the effective focal length (see [1.16](#)) to the focal length of the primary mirror. The final focal surface is curved inward (toward the sky) with radius of curvature related to the radii of curvature of the primary and secondary mirrors (R_{primary} and $R_{\text{secondary}}$, respectively) by

$$\frac{1}{R_{\text{focal}}} = \frac{1}{R_{\text{primary}}} + \frac{1}{R_{\text{secondary}}} \quad (1.18)$$

There are many variants of the classical Cassegrain design. For professional optical astronomy, Ritchey and Chrétien designed the most widely used variant in the 1910s. Both mirrors in a *Ritchey–Chrétien (R–C) reflector* are convex hyperboloids, a combination that eliminates low-order spherical and comatic aberration across a large field of view at the cost of increased field curvature.

For some applications, it is desirable to transfer the reflector optical beam to an instrument that is not mounted on the telescope but located in a fixed physical location where it will not flex mechanically as the telescope moves. For many years, the common solution was to use a flat tertiary (“third”) mirror to direct the telescope beam through an axis of rotation to the so-called *coudé focus* in a nearby room. For this to work, large telescope f-ratios were required to produce long enough focal lengths for the beam to reach from the primary mirror to the coudé instrument. When it was desirable for a telescope to have both Cassegrain and coudé foci, it was necessary to have separate secondary mirrors for each focus. In recent years, the introduction of alt-az mounts has allowed instruments that require gravity invariant environments to be located much closer to the telescope at the *Nasmyth focus*. Since Nasmyth and Cassegrain foci are separated from the primary mirror by the same physical distance (focal length), only one secondary is needed. This has become more important as primary and secondary mirrors grow in size and cost.

Many optical-infrared telescopes deploy *field correctors* in order to improve off-axis image quality and thereby increase the usable field of view. The highest priority is to decrease (flatten) field curvature but other optical aberrations can be minimized as well. The general topic of optical aberrations is discussed in more detail below. Field correctors are particularly important for instruments mounted at the telescope prime focus, that is, the focal point of the primary mirror. Prime focus systems are desirable because they provide the largest possible field of view

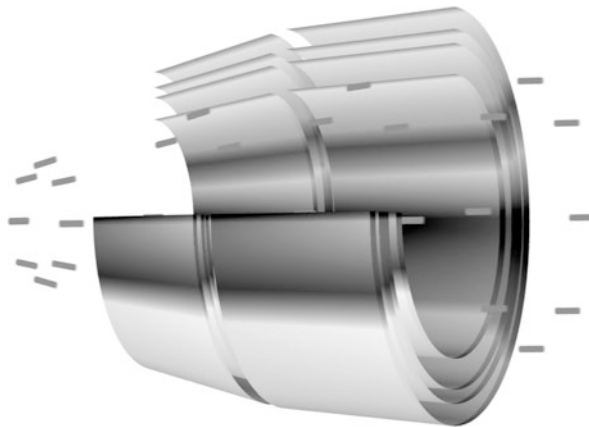
(aperture stop) and least loss of collected energy. These principles are exemplified by *Schmidt telescopes* that have spherical primary mirrors, aspheric field correctors located at that mirror's center of curvature, and a detector system at the mirror's prime focus. Additional field flattening optical elements are sometimes implemented in front of the detector system.

1.2.4 Grazing Incident Telescopes

At very short λ (very high energy), EMR arriving at nearly normal (small) incidence angles is not reflected or refracted, but is absorbed (by combination of photoelectric absorption and Compton scattering, depending on the material) or transmitted without any directional change. However, at small grazing angles (the complement of large incidence angles), EMR at x-ray wavelengths can be reflected and brought to a focus. The *critical grazing angle* is a function of λ and material properties. Heavier elements such as nickel, iridium, platinum, and gold are typically used at critical grazing angles in the range of $0.15\text{--}2^\circ$ for soft x-rays ($\lambda \sim 0.1\text{--}10\text{ nm}$, $E \sim 0.12\text{--}12\text{ keV}$). To establish and maintain appropriate optical shape, such films are deposited on rigid substrates. For most materials, x-ray reflectivity is not a smooth function of wavelength because strong absorption occurs when the photon energy corresponds to electron transition energies in the reflecting material. These absorption features are often called x-ray absorption edges.

Kirkpatrick and Baez (1948) were the first to implement grazing angle reflection techniques for x-ray imaging for microscopy. Their system used two sets of parabolic sheets with perpendicular axes of rotations. The first sheet focused incident light into a line (i.e., astigmatism), and the second set focused that line into a spot. Wolter (1952a, b) described a number of alternative approaches for x-ray microscopy that used pairs of parabolic and hyperbolic or parabolic and elliptical mirrors. Soon afterward, Giacconi and Rossi (1960) proposed an orbital x-ray imaging telescope that implemented a Wolter-type design.

Since the late 1970s, several space-based x-ray imaging telescopes have been launched. Most have used the so-called Wolter Type 1 design that consists of one or more coaxial and confocal parabolic and hyperbolic reflective shells (see [Fig. 1-9](#)). For any given parabolic-hyperbolic



■ Fig. 1-9

Grazing incidence telescope, Wolter Type 1, example (Illustration credit: NASA/SAO/Chandra X-Ray Observatory)

set, the collecting surface is only a small annulus. To increase collecting area, the number of shells is increased. Existing x-ray telescopes have achieved sub-arcsecond image sizes for point sources. The design and construction of grazing incidence telescopes that can focus higher energy x-rays into images is an area of current research interest.

1.3 Physical Optics and Image Formation

Although the basics of geometric optics were empirically established by 1800, image formation (and deformation) in a general sense was not well understood. Young's classic double-slit experiment changed all that by establishing the wave-like nature of light, which in turn laid the formation for a mathematical description of diffraction (interference) and hence ideal image formation. Actual images delivered by telescopes are distorted by imperfections in the optical elements and their relative alignments. For ground-based telescopes, atmospheric turbulence is the main source of image deformation at millimeter and shorter wavelengths.

1.3.1 Ideal Images

As an approximation, all distant luminous objects can be considered collections of point sources radiating spherical wave fronts. By the time those spherical waves reach Earth, they act like plane waves. In the nomenclature of geometric optics, astronomical objects exist in the far-field regime. When impeded by obstacles such as physical barriers or interfaces between media with different indices of refraction, these arriving plane waves bend. Such bending is the proximate cause of image formation by optical elements and devices and is known as diffraction or interference depending on the application.¹


Although diffraction was first described qualitatively by Grimaldi (1665), Fresnel (1815) used wave principles developed by Huygens (1690) (and vindicated by Young in the early 1800s) to develop the Huygens-Fresnel principle, a geometric theory for near-field diffraction, where incident waves are still spherical. The Fraunhofer diffraction equations were later derived for the case where a diffraction pattern is viewed a long way from the diffracting element or at the focal point of a positive optical element, that is, the case for all astronomical telescopes that form images. In the mid-twentieth century, the application of Fourier analysis to signal processing crossed over into classical optics (e.g., Duffieux 1946), bringing the appreciation that observed diffraction patterns can be described as the convolution of the Fourier transforms of the diffracting elements and the objects being observed.

Consider an idealized point source at great distance that is radiating monochromatic ($\Delta\lambda \rightarrow 0$) EMR that can be described as sinusoidal waves with form $x(t) = A \sin(2\pi ft + \varphi)$ where A , f , and φ are the amplitude, frequency, and phase, respectively. Such monochromatic EMR is said to be *coherent* (indeed, self-coherent), that is, all emitted EMR waves have phase and amplitude that are constant in time, and can therefore interact with itself to produce diffraction patterns.

¹Classically, interference occurs when waves overlap and combine to form resultant waves of greater or lower amplitude, while diffraction occurs when waves encounter obstacles that hinder wave propagation. From an image creation perspective, the distinction between interference and diffraction is essentially artificial.

In a vacuum, the image of that source produced by a single, round, perfect, and uniformly illuminated optical element is known as an Airy pattern (or disk) (from Airy 1835, the first mathematical description). In the nomenclature of Fourier optics, the Airy pattern is simply the squared modulus of the Fourier transform of a round, filled aperture (optical element):

$$I(\theta) = I_0 \left(\frac{2J_1(x)}{x} \right)^2, \quad x = kr \sin \theta \quad (1.19)$$

where I_0 is the maximum central intensity, J_1 is a first-kind Bessel function of order one, λ is the wavelength of incident light, $k (= 2\pi/\lambda)$ is the wavenumber, r is the circular aperture radius, and θ is the angle in radians between the central maximum and any other point in the pattern (see  Fig. 1-10). For any given λ , larger r leads to higher ideal spatial resolution as well as larger total (integrated) intensity since a larger area collects more photons.

Maximum central intensity I_0 is related to total power P_0 incident on the aperture by

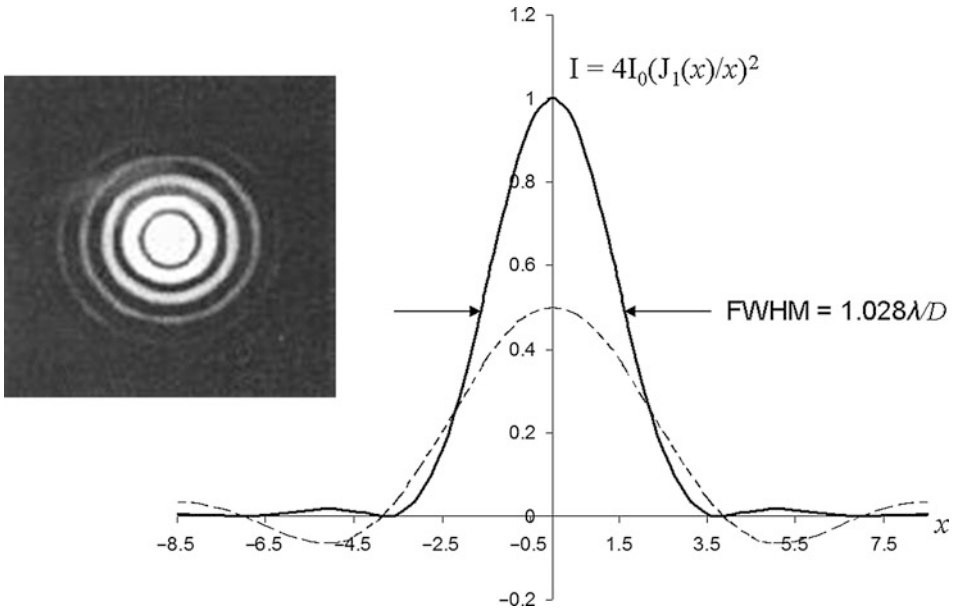
$$I_0 = \frac{P_0 \pi r^2}{\lambda^2 R^2} \quad (1.20)$$


where πr^2 is the aperture area and R is the distance from the aperture. Encircled power $P(\theta)$ is

$$P(\theta) = P_0 [1 - J_0^2(x) - J_1^2(x)], \quad x = kr \sin \theta \quad (1.21)$$

The angular width of the central maximum where intensity equals 50% of peak intensity I_0 is known as the *full width at half maximum (FWHM)*, that is,

$$FWHM = 1.028 \frac{\lambda}{D} = 0.51 \frac{\lambda}{r} \quad (1.22)$$



 Fig. 1-10

Airy pattern for circular aperture

where λ is the wavelength of the incident light and r and D are the circular aperture radius and diameter, respectively. Since plate scale (arcsec per linear distance) equals 206,265 divided by focal length, FWHM in linear units equals FWHM (radians) \times focal length. As focal length increases at constant aperture diameter D , images are magnified and collected energy is spread out over a large spatial area. So, images appear dimmer per unit area, and more time is necessary to record a fixed amount of energy per unit area in the image plane. Following Nyquist sampling theory, at least two spatial resolution elements must be contained within the delivered FWHM to avoid degrading spatial resolution during image detection and recording.

Peaks of intensity maxima (“fringes”) occur at $x \sim 5.136, 8.417, 11.617$, and so on while zero intensity minima (“dark rings”) occur at $x \sim 3.832, 7.016, 10.173$, and so on. The angular separation between the central maximum and the first zero is

$$\sin \theta_R = 1.22 \frac{\lambda}{D} = 0.61 \frac{\lambda}{r} \text{ (radians)} \quad (1.23)$$

where λ is the wavelength of the incident light and r and D are the circular aperture radius and diameter, respectively. This angular separation is known as the *Rayleigh criterion* or the “diffraction limit” because (in vacuum) two point sources can be separated (resolved) if the central maximum of one source lies at the position of this first minimum of the second source. Point sources are said to be fully resolved when the central cores of their Airy disks are completely separated by $2\theta_R$.

Obviously, many astronomical objects are extended, not point-like, and consist of many features of different shapes, sizes, and relative intensities. These features are only distinguishable if their intensity difference, their *contrast*, is large enough. Michelson (1927) defined contrast (also known as *visibility* or *modulation*) as

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \quad (1.24)$$

The concept of contrast can be simply illustrated by pairs of bright and dark lines (the comb function in Fourier analysis nomenclature). *Spatial frequency* is a measure of number of line pairs per unit interval. As spatial frequency increases, contrast decreases as the diffraction patterns from each bright line overlap more. The *contrast transfer coefficient* (CTC) is the ratio of the input contrast to the delivered contrast at a given spatial frequency, where $CTC = 1$ means input and output contrast are identically equal. The *contrast transfer function* (CTF) (also known as *modulation transfer function*, MTF) is essentially CTC as a function of spatial frequency. For an ideal system, $CTF = 1$ at all spatial frequencies. For a real system, CTF decreases as spatial frequency increases and eventually reaches zero at the *cutoff frequency*.

In an imaging system, the observed *point spread function* (PSF) is the Fourier transform of that system’s CTF (and vice versa). In optics design nomenclature, the PSF is the *impulse response* (or *function*) of a focused optical system, while in Fourier optics nomenclature, the PSF is the CTF (MTF) in the spatial frequency domain. For a point source, the first minimum (dark ring) in an Airy disk corresponds to zero contrast and a cutoff spatial frequency of $D/\lambda f$ where f is the focal length. In that case, spatial frequency ν and wavenumber k are related by $k = 2\pi\nu$ (see (1.19)). In short, as aperture diameter D increases, so does cutoff frequency, and therefore, a higher contrast (more detailed, higher fidelity, higher resolution) image is formed. The connection to the classic Rayleigh criterion is clear (see (1.23)). On the other hand, as observed EMR bandwidth $\Delta\lambda$ increases, EMR coherence decreases (especially in phase) and so does contrast. This effect is usually overwhelmed by the imperfections of real optical elements.

More generally, extended objects can be thought of as a two-dimensional collection of point sources where each point source produces its own PSF in the telescope focal plane. Since the EMR emitted by these point sources is incoherent (i.e., they have differing amplitude, frequency, or phase), the intensities of overlapping PSFs are additive. In short, the image of an extended object is the two-dimensional convolution of the spatial intensity of that object with the telescope PSF. Low spatial frequencies encode information about smooth (large spatial scale) intensity changes, such as general shape, orientation, and intensity gradients. High spatial frequencies encode information about sharp (small spatial scale) intensity changes such as edges.

1.3.2 Image Errors from Optical Elements

In a perfect system, an incident plane wave is transformed into an image wavefront that converges to an ideal diffraction pattern in the focal plane. The image wavefront is a two-dimensional map of phase relative to a reference surface, usually spherical. Real, imperfect systems introduce wavefront deviations at both high and low spatial frequencies. In physical terms, these deviations are phase changes, sometimes also called curvature, slope, or tilt errors. Classically, such wavefront errors are known as *optical aberrations* that arise from both imperfect individual optical elements and imperfect relative alignment of the optical elements. Optical aberrations are sometimes classified as intrinsic (created by design and/or manufacturing challenges) and induced (created by setup and/or operational errors). Relative to perfect images, aberrated images have reduced energy concentration, spatial resolution, and contrast.

To begin, consider three geometric issues that are not technically aberrations because the image wavefront is not deformed relative to the ideal case but merely misaligned. The first such geometric effect is *piston*, the mean value of the wavefront phase at the focal surface (really the exit pupil). The next two effects are *x* (sagittal) and *y* (tangential) *tilt* (tip-tilt), the average *x* and *y* slopes of the wavefront as it crosses the focal surface. Measuring and correcting piston and tip-tilt is critical to the performance of segmented mirrors and adaptive optics systems as described elsewhere in this volume.

The lowest order true aberration is *defocus*, the effect of forming an image before or after the optimal focal surface. It occurs when the imaging device (e.g., a camera detector) is placed in front or behind the effective focal point. The most common cause is incorrect separation between the primary and secondary mirrors. Defocus reduces image contrast (resolution). Fast (*f*-ratio ~ 3 or less) optical systems are very sensitive to defocus while slower optical systems are not.

The next five monochromatic ($\Delta\lambda \rightarrow 0$) aberrations are spherical aberration, coma, astigmatism, field curvature, and distortion. Spherical aberration is constant in magnitude everywhere in the focal plane but the magnitude of the other four aberrations increases with angular distance from the optical axis. In geometric optics, such aberrations are known as third-order errors because they can be described by third-order (or higher) terms of power series expansions of trigonometric functions as shown by von Seidel (1856). Hence, these five aberrations are also known as *Seidel aberrations*. In Fourier optics terms, the five Seidel aberrations correspond to low spatial frequency wavefront errors that affect the entire two-dimensional image.

Spherical optical elements focus off-axis and paraxial rays at different points along the optical axis. This effect is known as *spherical aberration* and can be either positive (wavefront bend too much) or negative (wavefront bend too little). Spherical aberration is more pronounced in short (fast) focal length systems. Multielement correctors that combine convex and concave

refractive elements (lens) in an appropriate fashion can be used to correct spherical aberration. The Hubble Space Telescope (HST) provides the most famous example of spherical aberration in the twentieth century. Due to a fabrication error, the HST primary mirror produces severe spherical aberration. Fortunately, it was possible to deploy optical systems to cancel that aberration, first through the Corrective Optics Space Telescope Axial Replacement (COSTAR) instrument and then later through correcting elements directly integrated into instruments deployed during later refurbishment missions.

Aspheric optical elements, such as parabolic or hyperbolic mirrors, can be used to minimize spherical aberration. However, such mirrors produce *comatic aberration* (*coma*). While point sources near the center of the field of view (on-axis) are well focused, point sources away from the center (off-axis) look wedge shaped, similar to comets. Coma results from a change in magnification with position that increases as focal ratio decreases (gets faster). It can be minimized by using multielement correctors or by appropriate combinations of primary and secondary mirrors. The Ritchey–Chrétien combination of a hyperbolic primary mirror with a hyperbolic secondary mirror is a widely used solution for minimizing coma (see [Sect. 1.2.3](#)).

Astigmatism is introduced when EMR traveling in two perpendicular planes (tangential and sagittal relative to the optical axis) is focused at two different focal distances. A common illustration is a plus sign where each arm is in focus at different distances from the imaging element.

While an input two-dimensional object plane is flat, the output image (focal) plane from a single optical element is curved. In essence, an overall curve (or change in radius) is added to the input wavefront. *Field curvature* increases as focal ratio decreases (becomes faster). To correct for field curvature, the imaging surface must be curved (e.g., the retina in a human eye) or the initial focal surface must be reimaged by other correcting optics. Without such correction, off-axis images are defocused relative to on-axis images.

Field distortion is the result of image scale (magnification) changing as a function of angular distance (radius) from the optical axis. Decreasing magnification with distance produces barrel distortion while increasing magnification with distance produces pincushion distortion.

Reflective optics create a single, common focal surface for all incident λ , while refractive optics in essence create an individual focal surface for each incident λ . This effect is known as *chromatic aberration*. Chromatic aberration is not *per se* a wavefront error, as each $\Delta\lambda$ can produce an unaberrated image.

All real optical elements induce one or more of the Seidel aberrations simply because fabrication errors are always nonzero to some extent. However, Seidel and other low spatial frequency aberrations can also be induced when the element shape is wrapped or twisted by gravitational slumping, incorrect mechanical support (e.g., pinching), or internal thermal gradients. Maintaining proper support and uniform temperatures is a significant design and operational challenge for large, massive telescope systems that are exposed to temporal temperature changes and must move to compensate for the rotation of the Earth.

Misalignment of optical elements is another source of low spatial frequency image aberrations. In a perfect system, all elements are centered on the optical axis, parallel to each other, and separated precisely so that each optical element images the focal surface of the preceding optical element. All three of these conditions can be violated individually or simultaneously. As general rules, centering errors (if small) cause optical aberrations to shift laterally on the focal surface without changing form, physical tilt errors increase (or create) coma, and separation (focus) errors reduce energy concentration as energy shifts from central core into diffraction rings and hence reduce image contrast. To maintain proper relative alignment, all optical elements must be supported within a common mechanical structure.

Real images suffer also from the effects of small-scale (high spatial frequency) wavefront errors. One common source of high spatial frequency error is diffraction from thin (sharp) mechanical obstructions such as secondary mirror support vanes and edges within segmented mirrors. Mirror surface roughness is another common source of high spatial frequency errors. The former error type produces sharp image features (e.g., diffraction “spikes”) while the latter type scatters energy into a low-intensity envelope (“noise floor”) throughout the more distinct diffraction pattern.

Additional diffuse background energy can also reach the focal surface without being captured and focused by the primary mirror, just from oblique stray rays entering from outside the entrance pupil. For small telescopes ($D \sim 2\text{-m}$ or smaller), a closed tube can prevent such stray light from reaching the primary. For large telescopes, mechanical barriers called baffles are often deployed to block stray EMR paths.

Obviously, a secondary mirror obstructs (“shadows”) the central area of the primary mirror and its effective collecting area. But this central obstruction also transforms the primary mirror from filled circle to an annulus, which is easily seen in out-of-focus images.² In turn, the system PSF is transformed from a classic Airy disk into the Fourier transform of an annulus, which has similar form with less total energy in the central core and more energy in the diffraction rings. Compared to a filled aperture, the effective CTF has relatively worse transfer at low spatial frequencies, comparable transfer at high spatial frequencies, and almost identical cutoff frequency. Thus, systems with relatively large obstructions maintain good image contrast, in the absence of other optical aberrations.

1.3.3 Image Errors from Atmospheric Turbulence

Before the wavefront from a celestial object reaches a ground-based telescope, it must pass through a column of terrestrial atmosphere that schematically consists of a series of turbulent cells, each with different density, temperature, and internal wind speed, and hence different indices of refraction. The exact mixture of cells in the line-of-sight changes rapidly with time. Turbulence as a function of altitude above a given location can be parameterized by the *index of refraction structure function*, $C_n^2(h)$. This structure function can be measured empirically but varies on both short and long timescales. Students of seeing often speak of two zones: the ground (boundary) layer right above the telescope location ($h \ll 2\text{ km}$) and the rest of the column where the jet stream ($h \sim 10\text{ km}$) is often the dominant source of turbulence. More sophisticated models use multiple layers. As the original wavefront passes through this turbulent column, it is deformed multiple times, reducing coherence and contrast. In effect, the wavefront passes through series of time-variable drifting lenses. The overall effect is to introduce *atmospheric seeing*.

Wavefront perturbations introduced by a given $C_n^2(h)$ can be predicted by models based on the atmospheric turbulence theory of Kolomogrov (1941). *Fried’s parameter* (r_0) is the spatial scale over which root-mean-square wavefront errors are less than 1 radian, the wavefront is unchanged (phase is constant, coherence is maintained), and it is possible to produce a diffraction-limited image. By definition, Fried’s parameter corresponds to the diameter of a circular aperture that will deliver the same image resolution an atmospheric column with a given $C_n^2(h)$. Telescopes with aperture diameter $D < r_0$ produce diffraction-limited images while

²The position and shape of the central hole in these out-of-focus “donut” images can be quite informative when trying to determine the presence and magnitude of the classic Seidel aberrations.

telescopes with $D > r_0$ produce so-called seeing limited images.³ As r_0 decreases, contrast decreases, that is, the atmospheric CTF has a smaller cutoff spatial frequency. The *isoplanatic angle* θ_{iso} is the observed angular size of the *isoplanatic patch* over which coherence is maintained. Fried's parameter r_0 and mean turbulence height \bar{h} are related to θ_{iso} by

$$\theta_{\text{iso}}(\text{arcsecs}) \sim \frac{r_0}{\bar{h}} \times 206,265 \quad (1.25)$$

The timescale between one wavefront deformation and another (and hence the timescale for coherence, contrast and seeing changes) is parameterized by τ_0 , the *coherence time*. The Fried parameter and coherence time are inversely related by the wind speed v in the dominant layer, that is, $\tau_0 = r_0/v$.

For plane waves, $C_n^2(h)$ and r_0 are related by

$$r_0 = \left[0.423 k^2 \sec \zeta \int C_N^2(h) dh \right]^{3/5} \quad (1.26)$$

where k is the wavenumber $2\pi/\lambda$ and ζ is the zenith distance (i.e., angle between zenith and optical axis of telescope). Since $r_0 \propto \lambda^{6/5}$, the effects of atmospheric seeing decrease with increasing wavelength and are negligible for radio telescopes operating at centimeter and longer wavelength.

Schematically, an object wavefront passes through a different turbulence column (isoplanatic patch) every time interval τ_0 . The dominant changes are piston (mean phase) and tilt (mean intensity centroid). The telescope transforms each modified wavefront into a separate image with slightly different aberrations at a slightly different position in the focal plane. Each of these instantaneous images is known as a *speckle*. The rapid position and intensity changes from speckle to speckle causes *scintillation*, which is further complicated by interference between deformed wavefronts that are not completely incoherent. Over $t > \tau_0$, for a point source, the intensities of these individual speckles are summed into a two-dimensional, pseudo-Gaussian intensity profile (*seeing disk*) with full width half maximum related approximately to r_0 by:

$$FWHM(\text{arcsec}) \approx \frac{25.1 \times \lambda(\mu\text{m})}{r_0(\text{cm})} \quad (1.27)$$

At optical wavelengths ($\lambda \sim 0.5 \mu\text{m}$), the best mountain sites (e.g., Mauna Kea, essentially all major Chilean observatories) have $r_0 \sim 20 \text{ cm}$ and $\tau_0 \sim 10 \text{ ms}$ for the corresponding to a free-air FWHM $\sim 0.63 \text{ arcsec}$ in the mean. It is not uncommon to find that the seeing-limited PSF actually has two Gaussian components: a semi-diffraction-limited "core" with FWHM $\sim \lambda/D$ and a "halo" with FWHM $\sim \lambda/r_0$ formed when r_0 is larger and smaller, respectively. This core-halo effect is more evident during periods of larger mean r_0 (better seeing).

Encoded in each speckle is an ideal image modified by aberrations introduced by atmospheric turbulence and the telescope itself. Since the 1960s, enormous effort by both civilian and military teams has gone into developing techniques to capture these images, characterize their inherent optical aberrations, and apply wavefront corrections in real time to produce near-diffraction-limited images over as large a field of view as possible. Broadly speaking, these

³For seeing limited telescopes, atmospheric turbulence effects dominate *delivered image quality*. Very pragmatically, this implies that ground-based telescope systems do not have to be perfect, they need merely to deliver images better than the site seeing. This is an extremely important concept toward minimizing design, construction, and operations costs.

efforts separate into two approaches. The first approach is to capture and record rapidly each individual speckle and then use *post facto* digital techniques to center each image to a common origin and sum their two-dimensional intensities, with or without additional filtering. Such *speckle imaging* has several variants. The second technique records each speckle (wavefront), immediately determines what wavefront errors are present, and then sends correction signals to a deformable surface to remove those errors in subsequent wavefronts. This technique is generically known as *adaptive optics* (AO). The topic of AO is revisited later in this chapter and elsewhere in this volume.

1.3.4 Parameterization of Image Quality and Wavefront Error

Achieved point source image quality and generalized wavefront quality can be parameterized in a number of ways and then compared to what a perfect system should produce.

For images of point sources, commonly used parameters include:

Full width half maximum (FWHM) – the PSF angular width of an observed point source at the position where the observed intensity is 50% of the peak observed intensity. The ideal, diffraction-limited value is given by (1.22). Observed FWHM is usually computed from a Gaussian or Lorentzian function fit to the observed PSF.

Encircled energy (EE) – really the radius of a circle that encompasses a certain fraction of light, for example, 80% (EE80) or 50% (EE50). For diffraction-limited images and circular apertures, EE80 is $1.38 \lambda/D$ (where D is the aperture diameter), that is, somewhat later than the angular size of the first Airy minimum (see 1.23). A larger EE80 means that energy has been redistributed from the central peak into other maxima. EE can be measured directly from real images of point sources and compared to EE calculated for ideal images.

Ellipticity (shape) – the radial behavior of image shape depends strongly on what aberration (or combination of aberrations) is dominant. Image shape is also a diagnostic of telescope tracking and wind buffeting rejection performance (discussed below). Since telescopes commonly have circular apertures, shape is often parameterized by ellipticity.

Optical path difference (OPD) – the path length difference between the ideal wavefront and the actual, aberrated wavefront. Systems with OPD less than $\lambda/4$ (quarter wave) are said to be diffraction-limited. As OPD increases, phase differences increase, interference decreases, and more energy is distributed in the other maxima (wings) of the diffraction pattern.

Strehl ratio (S) – the ratio of aberrated (achieved) PSF peak intensity to the diffraction-limited (theoretical) PSF peak intensity. If σ is the RMS wavefront deviation, then

$$S = e^{-(2\pi\sigma/\lambda)^2} \quad (1.28)$$

By definition, these metrics parameterize the quality of an in-focus image of a single point source at a specific position in the focal plane. Since most optical aberrations vary with radial distance from the optical axis, such on-axis measurements alone are not sufficient to characterize system performance completely. Two-dimensional approaches are necessary. While it is possible to obtain and analyze 2D images of a star fields, inherent intensity differences between stars and nonoptimal field coverage complicates *post facto* analysis.

The Hartmann (1900) test overcomes those problems by using an aperture plate to divide up the primary aperture into sub-apertures. When a single, bright star is observed out of focus, each sub-aperture produces a separate image with constant total intensity, known separation from other images, and aberrations as a function of radius from the optical axis. In fact, the bright star is observed at either side of the optimal focus. The observed spot diagrams can be compared to theoretical spot diagrams to determine the nature and magnitude of optical aberrations. Since the 1970s, more sophisticated wavefront characterization (“sensing”) techniques have been developed, including Shack-Hartmann, pyramid, and curvature. Contrary to the classic Hartmann test, the telescope pupil is split into separate sub-apertures or sub-images by optics in the focal plane. Since an aperture mask is not present, it becomes possible to measure wavefront errors in real time while also observing astronomical objects, by using a beam splitter to direct the telescope beam to both a wavefront sensor and a focal plane instrument. In turn, this enables real-time wavefront corrections to improve delivered image quality by driving active and adaptive optics systems, as discussed later.

Zernike polynomials (Zernike 1934) are often used to characterize (or reconstruct) observed wavefronts because they are orthogonal over a circle (the characteristic shape of telescopes optics) and three-dimensional (as are deformations from an ideal, unaberrated wavefront). The polynomials use polar coordinates ρ (normalized to 1 at the edge of the circle) and θ . They are defined by integer indices m (azimuthial frequency) and n (radial degree) where $m \leq n$ and $m - n$ is even. Each allowed shape $Z_n^m(\rho, \theta)$ is known as a *Zernike mode* with an index j . For values $m > 0$, each allowed (m, n) pair has two identical mode shapes except for rotation. The classic Seidel aberrations can be defined by various low-order Zernike modes, as listed in [Table 1-2](#) and shown in [Fig. 1-11](#). Remember that Zernike modes have three-dimensional shapes.

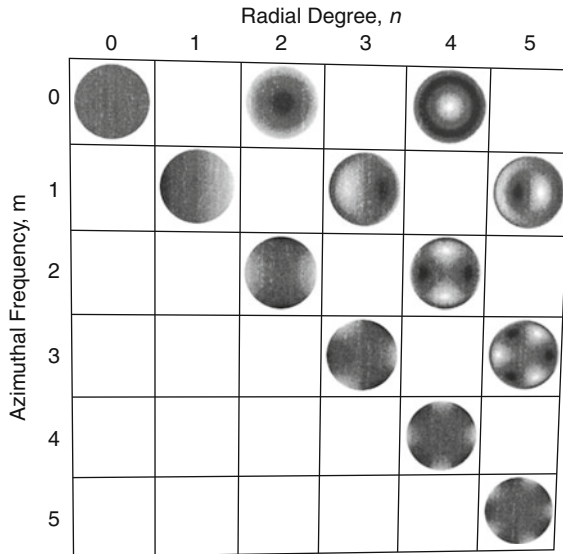
The total wavefront is characterized by

$$w(\rho, \theta) = \sum_j A_j Z_n^m(\rho, \theta) = \sum_j A_j Z_j(\rho, \theta) \quad (1.29)$$

where A_j is the relative contribution of each mode. The different summations only differ in assumed notation, not functional form. For wavefronts produced by static optical elements, values of A_j are fixed until some mechanical change is made (e.g., gravitational deformation, iterative fabrication). On the other hand, to model wavefronts modified by time-variable atmospheric turbulence, values of A_j must also vary every coherence time interval, τ_0 . The $A_j m$ variations are not completely independent of each other; rather, they are statistically coupled under the assumption that atmospheric turbulence is well modeled by a Komogoroff spectrum,

Table 1-2
Seidel aberration as function of Zernike mode

Seidel aberration	j	n	m
Piston	0	0	0
Tilt	1,2	1	1
Defocus (field curvature)	3	2	0
Astigmatism	4,5	2	2
Coma	6,7	3	1
Spherical aberration	8	4	0



■ Fig. 1-11
Zernike polynomials

that is,

$$\langle A_i A_j \rangle = C_{ij} \left(\frac{D}{r_0} \right)^{5/3} \quad (1.30)$$

where c_{ij} are the Noll coefficients (after Noll 1976), D is the aperture diameter, and r_0 is the Fried parameter.

Since sharp (high spatial frequency) errors are not well described by any kind of polynomial, they are usually parameterized by *structure functions* with form:

$$\Delta\phi(r) = \left\langle \left[\phi(r') - \phi(r' - r) \right]^2 \right\rangle \quad (1.31)$$

where $\Delta\phi(r)$ is the squared wavefront difference averaged over all separations $\Delta r = r' - r$. High spatial frequency (sharp) features are the domain of small separation. For ground-based telescopes, the goal is to fabricate optical surfaces with structure functions that do not degrade atmospheric seeing. This general rule implies that surface phase error is allowed to increase as separation increases as atmospheric seeing effects dominate at large separations, reducing fabrication costs. Similar growth in phase difference with separation is not acceptable for space-based telescopes.

1.3.5 Aperture Synthesis

In Young's double-slit experiment, observed fringe spacing (in radians) $\theta_f \sim \lambda/B$ where B is the slit separation while the fringe envelope is essentially the Fourier transform of the slit shape. If

round pinholes are used, that envelope is essentially an Airy pattern characterized by the diameter of the pinhole (see ● *Fig. 1-11*). Increasing slit separation reduces coherence and therefore fringe contrast (visibility, see ● 1.24). In Fourier terms, when zero contrast is reached, so has the cutoff spatial frequency and hence the maximum spatial resolution. In the late 1800s, Fizeau and Michelson realized independently that these principles could be inverted to measure the diameters of distant objects. In the latter case, Michelson derived the basic principles behind single pair (or rather single baseline) *spatial interferometry*, which he used most famously with Pease in 1920 to measure the angular size of a star beyond the Sun for the first time. Simply put, if coherence can be established and maintained between two telescopes with separation B , their received EMR can be combined to produce fringe patterns with separation $\sim \lambda/B$ contained within an envelope with width $\sim \lambda/D$ where D is the aperture diameter of the telescopes (assuming they have the same aperture size).

In essence, a single-filled aperture is a very large collection of single baselines. Each virtual baseline measures visibility at a given point in the two-dimensional spatial frequency (image) domain, called the *u-v plane* by interferometrists. In other words, the image produced by a filled aperture telescope is the superposition of fringe patterns from all baseline combinations in the telescope pupil. Delivered image contrast (quality) is degraded by a combination of atmospheric seeing and telescope intrinsic PSF (i.e., the intrinsic CTF with its finite cutoff spatial frequency). In the ideal case, the deconvolution of the telescope PSF and this two-dimensional visibility function produces an accurate representation of the observed object in the spatial (object) domain. If the primary mirror is blocked somehow (e.g., by a secondary), not only is total collected energy reduced, but spatial frequency information is lost, reducing image quality (contrast). For many interesting objects, especially ones with simple geometries, such losses are not critical.

Aperture synthesis (often synthesis imaging) emerged from these concepts, first at radio wavelengths in the 1950s and then at optical wavelengths in the 1980s. In essence, sparse arrays of telescopes are deployed to create an imaging system with high spatial resolution determined by the largest pair separation at the cost of incomplete, instantaneous spatial frequency coverage and a total collecting area only equal to the sum of the collecting areas of the individual telescopes. To increase image fidelity, more visibilities and more *u-v* points must be measured. Adding more telescopes to measure more visibilities simultaneously can accomplish that goal while also increasing total collected energy. In practice, all existing arrays use a different approach – they periodically change the geometric arrangement of the existing telescopes and/or let the Earth rotate to create different projected baselines. As new baselines emerge, visibilities can be measured at different *u-v* points. With patience, the *u-v* plane can be explored as thoroughly as desired for a given research problem.

Interferometers can only work if both amplitude and phase can be measured at each telescope and the signals from each pair combined coherently. In practice, this requires precise control of the optical path difference between any telescope pair. At radio wavelengths, this can be done electronically, sometimes waiting to combine (correlating) the signals until a later time if very large distances separate the telescopes. Radio interferometers working at centimeter and millimeter wavelengths have been very successful over the last 40 years.

The technical challenges of optical interferometry are much more daunting, primarily because the signals must be combined and measured in real time while the optical path difference must be controlled to submicron precision. Signal decoherence arising from atmospheric turbulence introduces a challenge not found in radio interferometers. The intricacies of optical interferometry are explored in more detail later in this volume.

1.4 Ancillary Systems

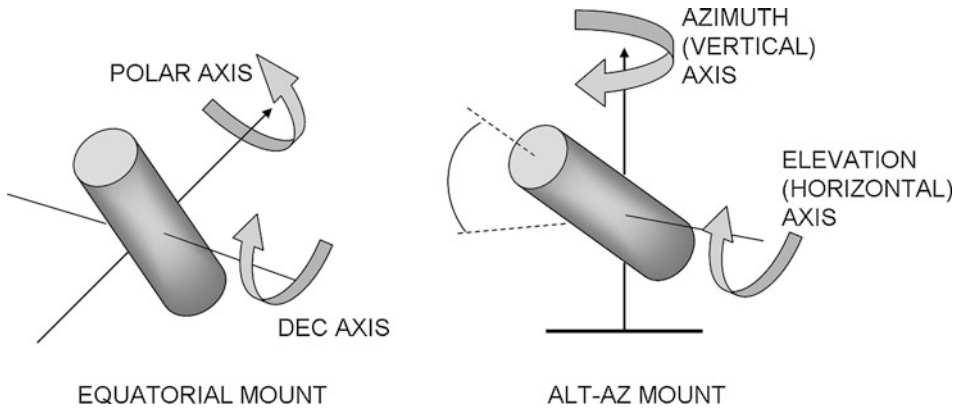
Modern telescopes are not merely static sets of high-precision optics – they are massive (100s of metric tons), highly dynamic, interconnected, automated systems of subsystems. Modern telescope systems must maintain the proper shapes and alignments of their optical components, point them toward the right point on the celestial sphere, compensate for Earth's rotation (or spacecraft orbital motion), reduce the impact of turbulence induced seeing effects, and protect themselves from inclement environmental conditions. In this section, some of these practicalities are discussed, with an emphasis on ground-based, nearly normal incidence optical reflectors. Aspects of radio telescopes and space-based systems are interweaved as a secondary theme.

Throughout this section, a common theme is minimize wavefront deformation to maximize image fidelity. It is a continual challenge during design, construction, and operation to identify potential sources of wavefront error and minimize them in a balanced way. As one illustrative example, at optical wavelengths, mirror surface accuracy requirements can be more relaxed for ground-based telescopes than space-based telescopes because ground-based delivered image quality is dominated by atmospheric seeing at those wavelengths. The nature and magnitude of all potential error sources can be captured in a delivered image quality *error budget*.

Although maximizing image fidelity is the primary goal for all telescopes, maximizing total system throughput of incident energy is an important secondary consideration. A telescope designed to achieve excellent image quality at the cost of highly attenuated energy transmission is not very attractive. The first step is to create the largest possible collecting area by maximizing aperture (primary mirror) size and minimizing obstructions between its surface and the energy source. Next is to minimize the number of reflections, transmissions, and obstructions between the primary mirror and recording device. Since no real mirror or lens is perfect, more optical elements means more energy loss and lower system throughput. Obstructions can both prevent energy transmission and scatter energy. The latter increases the observed background, resulting in decreased image contrast. Finally, each optical element must be tuned appropriately for the wavelength of interest. For mirrors, that means obtaining spectral smoothness and using reflective coatings appropriate for that wavelength (see [Section 1.2.1](#)). For lenses (used frequently, e.g., in field correctors), that means using high purity transmissive materials and multi-layer anti-reflection coatings. For optical-infrared ground-based telescopes, the last steps are to clean each air-exposed surface regularly (weekly to monthly) and re-coat them more infrequently (every several years).

1.4.1 Telescope Mounts and Motion Control

Whether on the ground or in space, telescopes must in general move to bring objects of interest (*targets* for short) into the telescope field of view (FOV) and then must continue to move to keep them in the FOV. The former process is known as *target acquisition* while the latter is called *target tracking*. These motions must be accomplished at some appropriate rate (angular change per unit time) and smoothness. All telescopes have a mechanical *telescope structure* that envelops the optical elements and allows telescopes to move as semi-rigid bodies. Most telescope structures are made from steel but substructures made from carbon-fiber-reinforced polymers (CFRP) are growing more common. Ground-based telescopes have low friction *bearings* that carry telescope weight and enable motion. To move ground-based telescopes, torque is



■ Fig. 1-12
Equatorial and altitude-azimuth (alt-az) mounts, schematic

applied by *drive systems* that consist of *motors* to generate the torque, *encoders* to track telescope position, and *servomechanisms* (servos for short) to control the motion through closed-loop feedback. As an ensemble, these structures and mechanisms are called the *telescope mount*. The mount is anchored to a concrete *pier* that in turn is anchored in bedrock.

While free-floating spacecraft do not require bearings or drive motors, they still need to acquire and track targets under servo control. Such orientation changes can be accomplished by thrusters, flywheels, and/or gyroscopes. Some spacecraft implement several solutions (e.g., Wilkinson Microwave Anisotropy Probe, WMAP, uses thrusters and gyroscopes), while some only use gyroscopes (e.g., Hubble Space Telescope). Position information can be provided both directly (e.g., Sun or star trackers) or indirectly (e.g., inertial motion tracking systems).

All telescopes must have at least one axis of motion, most have two axes, and space-based telescopes usually have three. Two coordinate systems are used to define ground-based telescope motion. In the *equatorial system*, one axis points along the Earth's axis of rotation (enabling motion in hour angle or right ascension) and the other points toward the celestial equator (enabling motion in declination).⁴ Various types of equatorial mounts have been implemented over time but they are impractical for very large modern optical and radio telescopes. Introduction of digital control technology enabled the use of the *altitude-azimuth system* (or *alt-az* for short) where one axis is orthogonal to the Earth's surface (enabling motion in azimuth) while the other axis is parallel to the Earth's surface (enabling motion in altitude). Because they are free-floating, space-based telescopes can move in three dimensions. The relative orientation of the telescope to itself and the Earth is known as its *attitude* (► Fig. 1-12).

During target acquisition (also known as *telescope pointing*), a mean target position must be converted from known (catalog) celestial coordinates (right ascension and declination) specific to a fixed date (epoch) to the desired physical telescope attitude. Various motion, atmospheric, and mechanical effects must be taken into account. For ground-based and orbiting telescopes, the dominant factor is the motion of the Earth – its annual revolution around the Sun (and the associated effects of parallax and aberration), its daily rotation, and long-duration changes in axis of rotation orientation relative to the celestial sphere (nutation and precession with periods

⁴Analogous to longitude and latitude on the Earth's spherical surface, right ascension and declination are used to specify position on the celestial sphere.

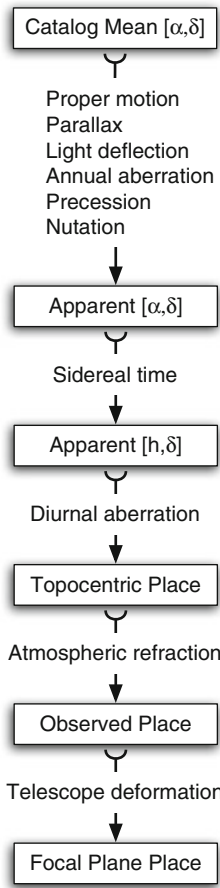


Fig. 1-13
Pointing model, transformation steps

of 18.6 and 26,000 years, respectively). Next, the apparent celestial motion of targets relative to their catalog position must be removed. For objects in the local solar neighborhood, this is called proper motion and it can be significant. For solar system objects, these motions can dominate the pointing solution! Since mean target position is given for a specific date and time, the time of target acquisition must be known precisely to adjust for terrestrial and celestial motion. For ground-based telescopes, two final steps are necessary: compensating for atmospheric refraction (that changes apparent position on the sky) and gravity-induced mechanical deformation (sag) of the telescope mount (that changes where the extrapolated optical axis “pierces” the celestial sphere) (● Fig. 1-13).

Generically, every modern telescope builds up a *pointing model* that accepts mean position and current time as input, produces desired telescope attitude as output, and then transmits that output to the servos that change telescope attitude so that the extrapolated optical axis reaches the desired position on the celestial sphere. All-sky RMS pointing precision (i.e., the achieved celestial position of the projected optical axis vs. the desired position) of ± 2 arcsec or better on timescales of minutes from point-to-point has been achieved by many modern telescope

mounts. Small field-of-view *acquisition cameras* are used to detect and correct small pointing errors during target acquisition.

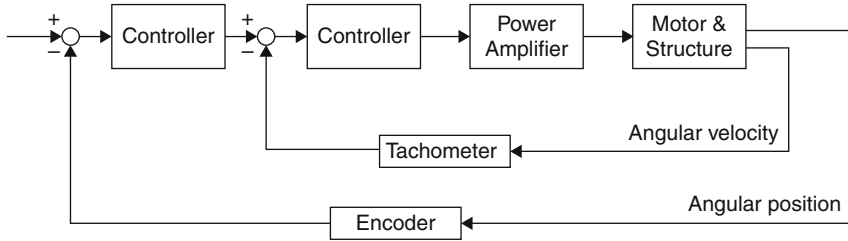
As the target acquisition ends, target tracking begins. Fundamentally, telescope motion while tracking should not degrade the image quality delivered by the convolution of atmospheric seeing and the native telescope PSF. Tracking has two components. The first component is *open loop tracking*, that is, smooth and continuous motion to compensate for Earth's rotation (on the ground) or motion through space (in orbit or free-floating). For equatorial mounts, the mount is rotated about the equatorial axis at the Earth's rotation rate, the sidereal rate.⁵ For alt-az mounts (and spacecraft), the pointing model is used continuously to determine desired position and send position commands to the servos at each axis. Open-loop tracking is never perfect. Sources of tracking errors include pointing model problems that produce inaccurate position updates and tracking trajectories, servo control problems that limit position update precision, and fundamental mechanical problems that prevent smooth enough motion. At radio wavelengths, these errors are often negligible because the native telescope PSF is large relative to tracking errors. Not so at millimeter and shorter wavelengths, where open-loop tracking is rarely precise enough to maintain image quality over long periods of times. Closed-loop tracking, that is, *auto-guiding* or optical feedback, overcomes these problems by continuously measuring the position of a star in the telescope FOV and transmitting small position updates to the mount servos to correct inaccuracies in open-loop tracking. Position updates on seconds (~1 Hz or slower) timescales are adequate. Faster updates to the mount servos can excite high-frequency mount vibrations (resonances) that will degrade image quality, defeating the primary purpose of tracking.

To move ground-based telescopes smoothly and accurately, low friction bearings are critical. While exact details vary between telescopes, in practice, there are two bearing types: mechanical and hydrostatic. The former are all variants of wheels moving on smooth surfaces (e.g., bogies on tracks, rollers on rails) while the latter injects a thin (10s of microns) film of oil between two flat surfaces. In essence, the entire telescope mass (often 100s of metric tons) floats on that film with such low friction that a single person can move the telescope simply by pushing on it.

Drive systems have three parts: electric motors that apply torques to move the telescopes, analog-to-digital encoders to determine and report telescope position, and servos to provide automated control. This combination is repeated in form for each axis of motion but usually with different details. Motors can be coupled to axes in three ways: gears (motor turns an axis connected to a gear that is mechanically coupled to the telescope axis, e.g., rack-and-pinion type), friction (motor turns an axis connected to a gear that is mechanically coupled to the telescope axis by friction alone), or direct (motor stator is constructed around the axis so that the axis effectively becomes the motor rotor). Encoders can also be coupled to axes in three ways: gears (encoder counts revolutions of geared shaft coupled to axis), friction (encoder counts revolutions of wheel coupled by friction to axis), and optical (encoder reads markings analogous to barcodes on axis). Encoders that use gears or friction can only determine a position relative to some starting point, while optical encoders can determine position absolutely because the axis markings are static and permanent.

The task of the main axis servos is to accept desired position updates during target acquisition and tracking and then move the telescope to that desired position as quickly and smoothly

⁵To observe solar system objects, it is necessary to drive telescopes at nonsidereal rates, that is, faster than rotation of Earth and on trajectories on angles with the celestial equator.



■ Fig. 1-14
Servomechanism, schematic

as possible without harming the motors (by demanding more torque than they can safely deliver) or the telescope components (by nonuniform “jitter” that rapidly shakes the telescope directly or induces uncontrollable vibrational resonances).⁶

❶ *Figure 1-14* illustrates a common, straightforward servo setup for motion control of one telescope axis.

1.4.2 Active Optics and Optics Control

For the most part, space-based telescopes operate in static environments. After launch, deployment, and commissioning, additional mechanical adjustments are rare.

Not so for ground-based telescopes, where time-variable gravity, thermal, and wind loads can and do cause mirror shape and alignment changes that introduce wavefront errors and degrade delivered image quality. Gravity and thermal effects produce smooth (low spatial frequency) changes on a timescale of minutes. Wind effects are more complicated since wind can induce significant spatially variable pressure changes on small timescales. To minimize these effects, each mirror must be embedded within mechanical designed to compensate for variable gravity, thermal, and wind loading by both passive and active means.

Radio telescopes must have large primary mirrors to provide enough collecting area to detect cosmic sources at these long wavelengths. But because wavelength is large (millimeters and larger), relatively large surface errors (micron to millimeter scale) can be tolerated. To compensate for variable gravity loading, many radio telescopes are constructed so that overall mechanical structure deforms under gravity in such a way that the primary mirror remains parabolic at all time. This causes small changes in the position of the primary focal point, so it is necessary for the instrument or reflector at prime focus to move in compensation. Surface changes induced by thermal and wind loading can be rendered negligible by passive mechanical means in most cases. Load compensation becomes more important as EMR wavelength decreases from meter to millimeter scales.

Operating at smaller wavelengths, the shape and alignment of mirrors in optical telescopes must be controlled much more precisely (nanometer scale). In the first telescopes, mirrors were supported passively – mechanical constraints were adjusted during construction and then only

⁶No real mechanical structure is infinitely stiff. Torque changes that are applied too quickly and/or too frequently on one axis can induce uncontrollable motions at a more distant point. Think of small shakes at one end of a thin piece of wood causing much larger motion at the other end.

readjusted occasionally later. As optical mirrors increased in diameter, they also decreased in mass density per unit volume, either by removing unnecessary material (“light weighting”) and/or becoming thinner. In effect, mirrors become more susceptible to all three kinds of loading and more sophisticated compensation systems had to be implemented. Generically, these mechanical systems are known as *mirror cells*.

First, consider gravity loading, usually separated into two components: axial (parallel to the optical axis) and lateral (perpendicular to the optical axis). It follows that optical mirror cells have axial and lateral support systems, each consisting of a number of load-bearing mechanical supports. For all mirror cells, the overall goal is to maintain mirror shape (in both dimensions) and alignment (centered on optical axis, tilted properly relative to other mirrors). While once completely static, these support systems have evolved to become highly dynamic. Early work introduced axial support systems with a small number (usually three) static supports (called *hard points*) to establish a known reference point and a larger number of dynamic supports that adjusted pressure applied to the mirror telescope altitude changed. Common examples include counterweighted (astatic) levers and air-filled bellows.

Since the 1990s, most new professional telescopes have implemented *active optics* systems where mirror support and alignment is continuously updated under closed-loop servo control. The support systems themselves consist of actuators of one sort or another that push or relax in the axial direction and push or pull in the lateral direction. Such systems have two basic control loops. The slow (or outer) loop uses static lookup tables to adjust mirror shape and alignment as a function of telescope elevation (telescope tilt relative to the gravity vector) and temperature (telescope length and hence mirror separation) on a timescale of minutes. *Lookup tables* are established by measuring delivered wavefront errors as a function of elevation and temperature, computing corrective force matrices, and storing those matrices for later use. It can be difficult to determine which mirror causes which wavefront error, so it is common to arbitrarily correct low-order aberrations (tip-tilt, focus) by tilting and/or moving the secondary mirror and high-order aberrations (e.g., the Seidel aberrations) by applying appropriate force changes to the primary mirror. It can also be difficult to separate atmospheric- and telescope system-induced wavefront errors. The common solution is to measure wavefront error as a function of telescope elevation and temperature repeatedly to average out atmospheric effects. Lookup table accuracy is greatly improved when such measurements are done on nights of good seeing (large r_0).

A second (inner) loop is often used to adjust mirror shape and alignment based on wavefront measurements made by a wavefront sensor located in the telescope focal plane (or fed by a pickoff mirror in the focal plane). Deviations from the ideal wavefront are determined and then actuator force updates are computed. Since the outer loop is still running, the inner loop is essentially applying corrections to small imperfections in the outer loop lookup table. Again, it can be difficult to separate atmospheric- and telescope-induced wavefront errors, so it is common to restrict inner loop updates to low-order corrections using the secondary mirror.

Thermal loading creates two kinds of wavefront error for ground-based optical-infrared telescopes. First, temperature gradients within mirrors induce torques that cause shape deformations relative to the ideal shape. Second, a temperature difference at the mirror/air interface can induce small-scale turbulence and hence phase errors. In concept, the solution to both problems is straightforward – maintain an isothermal mirror at the current air temperature. In practice, this is nontrivial since air temperature often changes faster than it is possible to change the temperature of a large mirror by forced cooling, even within the course of a night. These thermal effects are negligible for radio telescopes.

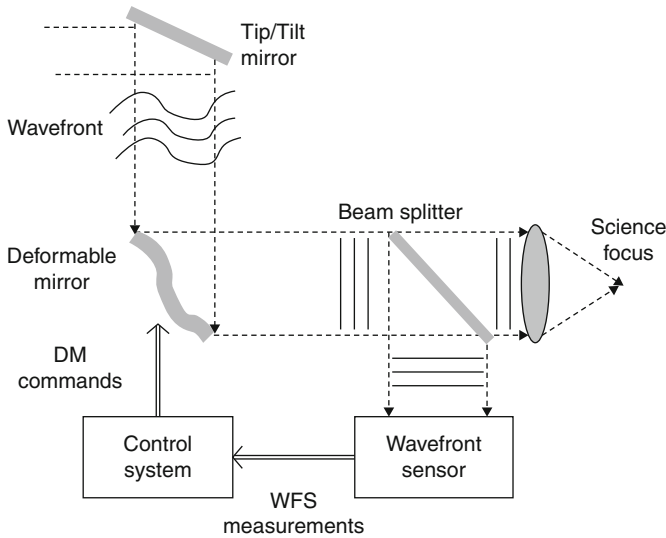
Wind loading has both static and dynamic components. The static component is a mean force related to the mean wind velocity. Typically, telescope mechanic and control systems are designed to maintain specified motion and optic support performance up to a given mean wind velocity, after which performance is allowed to degrade until some critical wind velocity is reached and the telescope must be stowed and/or enclosed to protect it. The dynamic component (known as *wind buffeting* or gusting) is much more problematic in two regards. First, wind buffeting can induce sudden telescope motions of small, random amplitude that cannot be stabilized by the main axis servo loops.⁷ Such motions introduce significant image blur (tip-tilt error) as the optical axis moves with the telescope. The availability of low-mass secondary mirrors capable for rapid tip-tilt motions enables a solution to this problem, at least up to some mean wind speed limit. If guide star flux is high enough (because star is bright or aperture is large enough), star trackers for closed-loop tracking systems can measure image centroids at very high rates (10s of Hz). Those centroids can be processed to determine mean image motion vectors on short (~ 2 Hz) and long (~ 0.2 Hz) intervals. The short timescale vectors are used to remove buffeting effects by rapidly changing secondary tip-tilt, while the long timescale vectors are used to correct open-loop tracking errors as discussed above. Second, wind buffeting can create differential pressure across the primary mirror surface, inducing shape changes that introduce sudden, rapidly changing wavefront errors. In some, but not all, wind buffeting regimes, these errors can be detected and corrected by adaptive optics systems (see next section). In many cases, the only recourse is to adjust enclosure openings to reduce wind loading, at the cost of reducing airflow and enclosure flushing. Finding the right balance between wind and thermal loading control is a topic of continuing active research.

Before leaving this topic, note that the largest optical primary mirrors in existence are segmented. The primary mirror support system must not only control the shape of each small (~ 1 -m segment) against gravity, thermal, and wind loading on a nanometer-size scale, the cell must also maintain the relative positions of the segments to each other to similar tolerance. In short, it is a very complex system of tens to hundreds of servo-controlled actuators. The intricacies of this challenge are discussed elsewhere in this volume.

1.4.3 Adaptive Optics and Turbulence Compensation

As discussed above, atmospheric turbulence creates incident wavefront deformations in the form of tilts (phase changes) at optical-infrared wavelengths, fundamentally limiting image quality (fidelity). Development of phase correction techniques generically known as *adaptive optics* (AO) has been a major effort since the late 1960s in the military world and later by civilian groups. Similar phase changes are introduced at submillimeter–millimeter wavelengths by changes in water vapor content integrated along the line of sight. Development of *phase calibration* techniques has been a major effort since the 1990s in support of the deployment of major new (sub-)millimeter interferometric systems such as ALMA.

⁷As primary mirrors become larger and/or more segmented, differential pressure changes across the mirror diameter caused by wind loading will introduce detectable wavefront errors and thus become a noticeable challenge. Compensation strategies for this challenge are a matter of active analysis and design for extremely large next-generation optical-infrared telescopes.

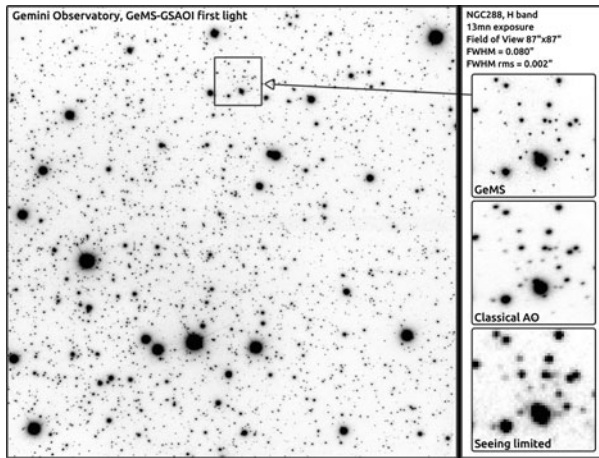


■ Fig. 1-15
Adaptive optics system, schematic

A schematic AO system is shown in ► Fig. 1-15. The EMR wavefront from a bright point source passes through the atmosphere and telescope, containing dominant phase errors from atmospheric turbulence and secondary errors caused by the telescope optics. It is then injected into the AO system where phase corrections are applied. The wavefront sensor (WFS) measures phases in the corrected wavefront, either directly or indirectly depending on WFS type. Measured wavefront information is passed to a control system that determines differences between measured and idea wavefront. The control system then issues wavefront correction information to the tip-tilt (woofer) and deformable mirrors (and in some instances to the telescope secondary mirror, thus establishing a connection to the active optics system). The tip-tilt mirror must be capable of motions on the scale of the incident EMR wavelength in order to compensate for low-order tip-tilt errors – it must have large stroke – but it can move as a rigid body. The deformable mirror must be capable of smaller motions – smaller strokes – at many points to compensate for high-order errors. By their nature, deformable mirrors are thin and flexible with many actuators to change their shape. Actuator number and spacing determine the highest spatial frequency that can be corrected. A beam splitter sends the corrected beam to the science instrument as well as back to the WFS and the cycle repeats. The correct-measure-correct cycle creates a closed-loop process.

This is the description of a *single conjugate AO (SCAO)* system using a natural guide star. For satisfactory phase correction, the natural guide star and target must lie within the same isoplanatic patch, (► 1.25) leading to two key limitations of SCAO systems. First, not all targets have a suitable (bright enough) guide star nearby. Second, as wavelength decreases, the isoplanatic patch decreases in size and phase changes occur faster. In principle, a constant level of phase correction can be maintained by increasing actuator density and update rate. In practice, the production of high-fidelity (near diffraction limited) images is limited to $\lambda > 1 \mu\text{m}$ for various technological reasons.

Driven by these limitations, various other AO approaches have been developed, including:



■ Fig. 1-16

Gemini Multi-conjugate adaptive optics System (GeMS) first-light image (Credit: Gemini Observatory/AURA)

Laser guide star AO (LGS-AO) – a laser tuned to the Na D resonance line ($\lambda \sim 589.3 \text{ nm}$) can create an artificial star in the atmospheric Na layer at altitude $\sim 90 \text{ km}$, essentially creating a guide star next to any target. Since tip-tilt errors are introduced into the artificial star when the laser beam passes upward through turbulence to the Na layer, a natural guide star must be observed to correct tip-tilt errors delivered to the science instrument. However, natural guide stars for LGS-AO systems can be much fainter than required for SCAO systems because only tip-tilt (position or centroid) information must be measured and because their contrast (energy concentration) is increased by the LGS-AO correction. Stars suitable for LGS-AO tip-tilt correction have high enough surface density on the celestial sphere to make LGS-AO practical, at least for 8 m and larger aperture telescopes.

Multi-conjugate AO (MCAO) – by using multiple natural or artificial guide stars (or sources), near diffraction-limited correction can be achieved over a much larger area than the natural isoplanatic patch. The European Southern Observatory (ESO), National Solar Observatory (NSO), and Gemini Observatory (GO) have all made recent impressive progress (see [Fig. 1-16](#)).

Ground-layer AO (GLAO) – partial phase correction over a FOV of several arcmin can be achieved by sensing phase errors created by low-altitude (ground-layer) turbulence less 2 km above the telescope, providing significant seeing improvements at $\lambda < 1 \mu\text{m}$. While multiple natural guide stars can be used, performance is best when a low-altitude ($\sim 10 \text{ km}$) laser guide star is used as well. Optical and UV lasers can be used to illuminate molecules and dust at such low altitude and create (via Rayleigh scattering) artificial stars. Such Rayleigh systems are much less expensive and complicated than Na layer systems. The Isaac Newton Group (ING) observatory, Steward Observatory (in partnership with the MMT Observatory), and the National Optical Astronomy Observatory (NOAO) (in partnership with the Southern Observatory for Astronomical Research, SOAR) have all deployed GLAO systems in recent years.

AO technology is advancing rapidly at present time. One of the most promising developments is the development of thin shell, adaptive secondary mirrors by groups at Steward and Arcetri Observatories (among others). Such secondary mirrors can move low-order and high-order correction “up stream,” significantly improving system energy throughput for SCAO, LGS-AO, and MCAO systems as well as enabling large FOV for GLAO systems.

Imaging synthesis arrays rely on the assumption that incident EMR has the same phase at all antennas (telescopes) in the array or that phase differences can be corrected in some fashion. Significant differences in received phase between array antennas are equivalent to phase errors between the emitted and received wavefronts for filled aperture optical-infrared telescopes. In both instances, final image quality (fidelity or contrast) is degraded.

At centimeter wavelengths, phase differences between antennas are dominated by instrumentation effects that vary slowly (timescale of minutes) that can be calibrated via observations of sources with known phase. At (sub-)millimeter wavelengths, phase differences are dominated by $C_n^2(h)$ variations caused by differences in the integrated water content along the line of sight of each telescope in the array that occur on timescales of seconds to tens of seconds. Two methods can be used to detect and correct these variations. During *fast switching*, each telescope switches between observations of a target source of unknown phase and a calibrator source of known phase at intervals of tens of seconds. Using *water vapor radiometry*, the line-of-sight water vapor content is measured in near real time at every telescope. From these measurements, phase differences are inferred and corrections are applied immediately or during *post facto* data processing. At optical wavelengths, phase differences are dominated by $C_n^2(h)$ variations caused by difference in atmospheric turbulence along the line of sight of each telescope on sub-second timescales. Phase correction techniques at these wavelengths use systems analogous to natural guide star AO systems to detect and correct phase differences.

This whole topic is irrelevant for space-based telescopes!

1.4.4 Enclosures and Protection from Inclement Environmental Conditions

Most telescopes need some kind of protection from their local environment. For space-based telescopes, the key challenges are maintaining proper temperature despite solar thermal loading and preventing damage to on-board detectors from over-illumination by bright objects (Sun, Earth, Moon). While various subsystems (e.g., detectors, control systems) are enclosed, the performance of ground-based radio telescopes is essentially immune to local environmental conditions and therefore they are typically left exposed to the elements. On the other hand, ground-based optical-infrared telescopes are always surrounded by enclosures that can be completely sealed during the day or under inclement weather conditions (e.g., high wind, precipitation) and opened during the night to allow cosmic EMR to reach the telescope and to allow air within the enclosure to mix with external air to achieve thermal equilibrium.

The classic enclosure is a hemispheric dome with a closeable slit that rotates on top of a tall cylindrical building (► [Fig. 1-17](#)). As initially conceived, such enclosures allow very little air transfer between the enclosure interior and exterior, even when the slit is open at night. By the 1970s, it was widely appreciated that telescopes surrounded by such enclosures were producing much worse images than the natural (free-air) atmospheric seeing. Although originally attributed to nonisothermal, turbulent air within the enclosure, eventually the dominant source of so-called dome seeing was found to be turbulent air near the surface



■ Fig. 1-17

NOAO Cerro Tololo Inter-American Observatory Blanco 4-m telescope (Credit: Roger Smith/NOAO/AURA/NSF)

of the primary mirror caused by temperature differences between the mirror surface and the surrounding air.

Since the 1980s, optical-infrared enclosure designs have strived to achieve three goals: minimize temperature differences at the primary mirror surface/air interface, minimize wind buffeting on telescope, and minimize cost. To achieve these goals, enclosure design must be tightly coupled with the design of other telescope systems. The solution to the mirror/air temperature interface problems has multiple components: minimize heat release by all telescope systems into the enclosure, actively cool the enclosure during the day when the enclosure is closed, actively manage mirror temperature at day and night, create many dome openings so that the air volume in the enclosure is flushed multiple times per hour under mean wind conditions at night, and design vent systems so that a laminar airflow is established across the primary mirror surface when the enclosure vents are open. Reducing the effect of wind buffeting involves a combination of variable venting as a function of wind speed, wind screen systems to reduce or eliminate wind pressure on telescope components with significant surface area, and active secondary mirrors that can remove tip-tilt errors produced by residual buffeting (see above). Finally, minimizing cost is related to minimizing enclosure size (and hence required raw material, such as steel) and mechanism count. Minimizing size has the added value of minimizing enclosed air volume, which enables more efficient air flushing and smoother airflow. While the next generation of extremely large telescopes all share these enclosure design goals, each has produced a different design concept, as presented later in this volume.

1.5 Looking Forward

In the 400 years since Galileo turned his telescope toward the sky and ushered in a revolution, telescope designs have evolved dramatically to adapt to new EMR windows and new environments, such as near-Earth space. In essence, all telescope builders must address several fundamental issues:

1. How to achieve a large collecting aperture of the required optical performance
2. How to move such a massive machine in an optimal fashion
3. How to control a system of subsystems in an optimal fashion
4. How to protect that system from inclement environmental conditions without reducing performance
5. How to maximize delivered image quality
6. How to minimize life-cycle costs

These topics and more are discussed by leaders in the field throughout the rest of this volume.

References

Selected References (Books)

- Anderson, T., & Enmark, A. 2011, *Integrated Modeling of Telescopes* (1st ed.; New York: Springer)
- Bely, P. ed., 2003, *The Design and Construction of Large Optical Telescopes*, (1st ed.; New York: Springer)
- Born, M., & Wolf, E. 1997, *Principles of Optics* (6th ed.; Cambridge: Cambridge University Press)
- Bracewell, R. N. 1999, *The Fourier Transform and Its Applications* (3rd ed.; New York: McGraw-Hill)
- Bracewell, R. N. 2004, *Fourier Analysis and Imaging* (1st ed.; New York: Springer)
- Cassegrain, L. 1672, in Letter, ed. M. de Bercé (Chartes, France), *Recueil des mémoires et conférences concernant les arts et les sciences* (trans: *Journal of memoirs and conferences concerning arts and sciences*), 1672 April 25
- Cavalieri, B. *Lo specchio ustorio: ovvero, Trattato delle sezioni coniche, et alcuni loro mirabili effetti intorno al lume, caldo, freddo, suono, e moto ancora* (trans: *the burning mirror, or a treatise on conic sections*) (Bologna, Italy)
- Descartes, R. 1637, *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences* (trans: *Discourse on the Method of Rightly Conducting One's Reason and of Seeking Truth in the Sciences*) (Leiden, Netherlands)
- Dorf, R. C., & Bishop, R. H. 2010, *Modern Control Systems* (12th ed.; Upper Saddle River, NJ: Prentice Hall)
- Duffieux, P. M. 1946, *L'intégral de Fourier et ses Applications à l'Optique* (Rennes: Société Anonyme des Imprimeries Oberthur)
- Duffieux, P. M., 1983, *The Fourier Transform and Its Applications to Optics* (2nd ed.; New York: Wiley)
- Florence, R. 1994, *The Perfect Machine: Building the Palomar Telescope* (1st ed.; New York: Harper-Collins)
- Galilei, G. 1610, *Sidereus Nuncius* (trans: *Starry Messenger*) (Venice, Italy)
- Glindemann, A. 2011, *Principles of Stellar Interferometry* (1st ed.; Heidelberg: Springer)
- Goodman, J. 2004, *Introduction to Fourier Optics* (3rd ed.; Greenwood Village, CO: Roberts and Company)
- Gregory, J. 1663, *Optica Promota* (trans: *The Advance of Optics*) (London, England)
- Grimaldi, F. M. 1665, *Physico mathesis de lumine, coloribus, et iride, aliisque annexis libri duo* (Bologna, Italy)
- Hardy, J. W. 1998, *Adaptive Optics for Astronomical Telescopes* (1st ed.; New York: Oxford University Press)
- Huygens, C. 1690, *Traité de la lumiere* (Leiden, Netherlands)
- Kepler, J. 1611, *Dioptrice*
- Malacara, D. (editor), 2007, *Optical Shop Testing*, (3rd ed.; Hoboken: Wiley)

- McLean, I. S. 2008, *Electronic Imaging in Astronomy: Detectors and Instrumentation* (2nd ed.; New York: Springer, Chichester, UK: published in association with Praxis)
- Mersenne, M. 1636, *L'Harmonie universelle* (trans: the harmonious universe) (Paris, France)
- Michelson, A. A. 1927, *Studies in Optics*, (1st ed.; Chicago: The University of Chicago Press) (also available 1995 from Dover)
- Newton, I. 1704, *Opticks, or, a treatise of the reflexions, refractions, inflexions and colours of light: also two treatises of the species and magnitude of curvilinear figures* (London, England)
- Schroeder, D. J. 1999, *Astronomical Optics* (2nd ed.; San Diego: Academic)
- Tyson, R. K. 2010, *Principles of Adaptive Optics* (3rd ed.; Boca Raton, FL: CRC)
- Rutten, H. G. J., & Van Venrooij, M. A. M. 1988, in *Telescopes Optics: Evaluation and Design*, ed. R. Berry (1st ed.; Richmond, VA: Willman-Bell)
- Wilson, R. N. 2007, *Reflecting Telescope Optics I: Basic Design Theory and its Historical Development* (2nd ed.; New York: Springer)
- Wilson, R. N. 2002, *Reflecting Telescope Optics II Reflecting Telescope Optics II: Manufacture, Testing, Alignment, Modern Techniques* (Corrected edition, New York: Springer)
- Wilson, T. L., Rohlf, K., Hüttenmeister, S. 2009, *Tools of Radio Astronomy* (5th ed.; Berlin: Springer-Verlag)

Selected Technical/Semi-technical References (Journal Articles)

- Airy, G. B. 1835, On the diffraction of an object-glass with circular aperture. *Trans. Camb. Philos. Soc.* 5, 283–291
- Aschenbach, B. 1985, X-ray telescopes. *Rep. Prog. Phys.*, 48, 579
- Einstein, A. 1905, Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt (trans: On a Heuristic Point of View about the Creation and Conversion of Light). *Ann. Phys.* 17(6), 132–148
- Fried, D. L. 1966, Optical resolution through a randomly inhomogeneous medium for very long and very short exposures. *J. Opt. Soc. Am.* 56, 1372
- Fresnel, A.-J. 1816, Mémoire sur la diffraction de la lumière $i_2^{1/2}$ (trans: Memoir on the diffraction of light $i_2^{1/2}$). *Ann. Chem. Phys.* 2nd series, 1, 239–281
- Frizeau, H. 1868, Prix Borodin: rapport sur le concours de l'année 1867. *Comptes Rendus de l'Academie des Sciences*, 66, 932
- Giacconi, R., & Rossi, B. 1960, A 'Telescope' for soft X-ray astronomy. *J. Geophys. Res.*, 65, 773
- Gorenstein, P. 2010, Focusing X-ray optics for astronomy. *X-Ray Opt. Instrum.*, Article ID 109740
- Hanbury Brown, R., & Twiss, R. Q. 1956, A test of a new type of stellar interferometer on Sirius. *Nature*, 178, 1046
- Hartmann, J. 1900, Bemerkungen über den Bau und die Justierung von Spektrographen (trans: Remarks on the construction and calibration of spectrographs), *Zt. Instrumentenk.* 20, 47
- Heisenberg, W. 1925, Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen (trans: Quantum theoretical interpretation of kinematic and mechanical relations). *Z. Phys.* 33(1), 879–893
- Kirkpatrick, P., & Baez, A. V. 1948, Formation of optical images by x-rays. *J. Opt. Soc. Am.*, 38, 766
- Kolmogorov, A. N. 1941, The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *C. R. (Dok.) Acad. Sci. URSS*, 30, 301
- Kolmogorov, A. N. 1941, Dissipation of energy in the locally isotropic turbulence. *C. R. (Dok.) Acad. Sci. URSS*, 32, 16
- Labeyrie, A. 1970, Attainment of diffraction limited resolution in large telescopes by Fourier analyzing speckle patterns in stars images. *Astron. Astrophys.*, 6, 85
- Mahajan, V. V. 1994, Zernike annular polynomials and optical aberrations of systems with annular pupils. *J. Opt. Soc. Am.*, 33, 8125
- Maxwell, J. C. 1861, On physical lines of force (in 4 parts). *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 21 and 23. (London, England)
- Meinel, A. B., 1965, Introduction to the design of astronomical telescopes, Technical Report No. 1, Optical Sciences Center (Tucson: University of Arizona)
- Michelson, A. A., & Pease, F. G. 1920, Measurement of the diameter of α Orionis with the interferometer. *ApJ*, 53, 249
- Monnier, J. D. 2003, Optical interferometry in astronomy. *Rep. Prog. Phys.*, 66, 789
- Noll, R. J. 1976, Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.*, 66, 207

- Planck, M. 1901, Über das Gesetz der Energieverteilung im Normalspektrum (trans: On the Law of Distribution of Energy in the Normal Spectrum). *Ann. Phys.* 4, 553
- Quirrenbach, A. 2001, Optical interferometry, *Ann. Rev. Astron. Astrophys.*, 39, 353
- Roddier, F. 1981, The effects of atmospheric turbulence in optical astronomy, in *Progress in Optics XIX*, ed. E. Wolff (Amsterdam: North-Holland), 281
- Roddier, N. 1990, Atmospheric wavefront simulation using Zernike polynomials. *Opt. Eng.* 29, 1174
- Seidel, L. 1856, Zur Dioptrik. Über die Entwicklung der Glieder 3ter Ordnung welche den Weg eines ausserhalb der Ebene der Axe gelegene Lichtstrahles durch ein System brechender Medien bestimmen (non-literal trans: Discussion of the development of third-order terms for off-axis light that passes through a system of refracting media), *Astro. Nach.*, 43, 289
- Snellius, W. 1621, unpublished (and now lost) manuscript, reported by Vossius and Huygens in later works.
- Wolter, H. 1952a, Spiegelsystem streifenden Einfall als abbildende Optiken für Röntgenstrahlen (trans: Glancing Incidence Mirror Systems as Imaging Optics for X-rays), *Ann. Phys.*, 10, 94
- Wolter, H. 1952b, Verallgemeinerte Schwarzschildsche Spiegelsysteme streifender Reflexion als Optiken für Röntgenstrahlen (trans: A Generalized Schwarzschild Mirror Systems For Use at Glancing Incidence for X-ray Imaging), *Ann. Phys.*, 10, 286
- Wyant, J. C. & Creath, K., 1992, Basic wavefront aberration theory for optical metrology, in *Applied Optics and Optical Engineering*, ed. J. C. Wyant, & R. R. Shannon (Orlando: Academic), 2
- Zernike, F. 1934, Beugungstheorie des Schneidverfahrens und seiner verbesserten Form, der Phasenkonstrastmethode. *Physica*, 1, 689
- Zernike, F. 1938, The concept of degree of coherence and its applications to optical problems. *Physica*, 5, 875

Selected Technical or Semi-technical Publications (Web-Based)

- Meister, D. 2010, Wavefront Aberrations and Spectacle Lenses, Parts 1 and 2, *Dispensing Optics Journal*, http://www.opticampus.com/files/wavefront_aberrations_and_spectacle_lenses.pdf
- Quirrenbach, A. 2002, The Effects of Atmospheric Turbulence on Astronomical Observations, Center for Adaptive Optics Summer School, http://cfao.ucolick.org/aosummer/book/pdf/3.1_quirrenbach.pdf
- Sacek, V. 2006 (with continuing updates), *Amateur Telescope Optics*, <http://www.telescope-optics.net/>
- Tripoli, N. 2003, The Zernike polynomials [http://www.optikon.com/en/articles/keratron_023/media/TheAberrometers_2003_Tripoli%20\(Zernike%20Polynomials\).pdf](http://www.optikon.com/en/articles/keratron_023/media/TheAberrometers_2003_Tripoli%20(Zernike%20Polynomials).pdf)
- Wallace, P. 2012, *Telescope Pointing*, <http://www.tpsoft.demon.co.uk/pointing.htm>

SPIE Conference Proceedings

SPIE (formerly – Society of Photo-Optical Instrumentation Engineers, now – The International Society for Optical Engineering) organizes major symposia on all aspects of observatories, telescopes, and instrumentation. Proceedings from the bi-annual *Astronomical Telescopes and Instrumentation* (1996 – 2012) symposia provide a comprehensive, in-depth overview of all aspects of a field in constant evolution.

2 Robotic and Survey Telescopes

Przemysław Woźniak

Los Alamos National Laboratory, Los Alamos, NM, USA

1	<i>Introduction</i>	45
2	<i>Design Considerations for Autonomous and Survey Telescopes</i>	47
2.1	Modular Design and Logistics	47
2.2	Optical Design	48
2.3	Telescope Mounts and Fast Slewing	51
2.4	Detectors	51
2.5	Weather Protection	54
2.6	Control Systems, Data Acquisition, and Processing	54
3	<i>Diversity of Survey and Follow-Up Instruments</i>	56
3.1	All-Sky Monitors	56
3.1.1	CONCAM	57
3.1.2	RAPTOR-Q	58
3.2	Mid Range Robotic Telescopes	60
3.2.1	ASAS	60
3.2.2	HAT	61
3.2.3	WASP and SuperWASP	61
3.2.4	ROTSE	62
3.2.5	RAPTOR/Thinking Telescopes	63
3.2.6	“Pi of the Sky”	65
3.2.7	MASTER	66
3.2.8	KAIT	67
3.2.9	PAIRITEL	68
3.2.10	REM, TORTORA, and MegaTORTORA	68
3.3	Dedicated Survey Telescopes	70
3.3.1	Microlensing Searches: OGLE and MOA	70
3.3.2	Near-Infrared Surveys: 2MASS, VISTA, and SASIR	71
3.3.3	Moving Object Searches: Spacewatch, LINEAR, and CSS	72
3.3.4	Spectroscopic Surveys	73
3.4	Large Robotic Telescopes	74
3.4.1	Liverpool Telescope	74
3.4.2	GROND	77
3.5	Large-Scale Surveys	78
3.5.1	SDSS	78
3.5.2	Palomar Transient Factory	78

3.5.3	Pan-STARRS	79
3.5.4	LSST	81
4	<i>Telescope Networking</i>	83
4.1	Communication Protocols and Virtual Observatory Standards	85
4.2	Real-Time Event Brokers	87
4.3	Automated Transient Detection and Classification	87
4.4	Follow-Up Optimization	88
4.5	Global Telescope Networks and Deployable Computing	89
5	<i>Science with Robotic and Survey Telescopes</i>	90
5.1	Cosmic Explosions	90
5.2	Solar System Science	91
5.3	Extrasolar Planets	92
5.4	Variable Stars	92
6	<i>Conclusions and Future Outlook</i>	94
	<i>References</i>	95

Abstract: Robotic telescopes are revolutionizing the way astronomers collect their data and conduct sky surveys. This chapter begins with a discussion of principles that guide the process of designing, constructing, and operating telescopes and observatories that offer a varying degree of automation, from instruments remotely controlled by observers to fully autonomous systems requiring no human supervision during their normal operations. Emphasis is placed on design trade-offs involved in building end-to-end systems intended for a wide range of science applications. The second part of the chapter contains descriptions of several projects and instruments, both existing and currently under development. It is an attempt to provide a representative selection of actual systems that illustrates state of the art in technology, as well as important ideas and milestones in the development of the field. The list of presented instruments spans the full range in size starting from small all-sky monitors, through midrange robotic and survey telescopes, and finishing with large robotic instruments and surveys. Explosive growth of telescope networking is enabling entirely new modes of interaction between the survey and follow-up observing. Increasing importance of standardized communication protocols and software is stressed. These developments are driven by the fusion of robotic telescope hardware, massive storage and databases, real-time knowledge extraction, and data cross-correlation on a global scale. The chapter concludes with examples of major science results enabled by these new technologies and future prospects.

Keywords: Asteroids, Astronomical databases: miscellaneous, Catalogs, Gamma rays: bursts, Gravitational lensing, Instrumentation: detectors, Instrumentation: miscellaneous, Methods: data analysis, Methods: observational, Minor planets, Miscellaneous, Planetary systems, Solar system: general, Spectrographs, Standards, Stars: variables, Statistical, Supernovae, Surveys, Telescopes

1 Introduction

Sky surveys play a crucial role in astrophysical research. A fundamental problem in astronomy, as in any observational science, is our inability to perform controlled experiments on most objects of interest. We only have one universe, given to us as the observable slice through vast expanses of space and time. Therefore, we need to thoroughly map its content and build statistical ensembles in order to learn about the inner workings of astrophysical objects, understand where they came from, and predict what their future fate will be. With the help of systematic sky surveys, we can also build uniform samples of rare objects and transient phenomena that occur at random times and locations and are otherwise hard to find. Much of the progress in observational astrophysics is made possible by a healthy interplay between the surveys and follow-up observations. The role of surveys is not only to build large statistical samples and initially characterize as many sources as possible but also to identify the most interesting ones for detailed follow-up studies that often require the use of large telescopes and specialized instruments such as spectrographs and polarimeters. Time-domain work creates its own challenges such as the need for prompt response. The focus of this chapter is on optical telescopes and supporting technologies that enable massive photometric sky surveys and highly automated follow-up programs. Although the following discussion is limited to professional efforts, it should be recognized that robotic astronomy benefits from a strong involvement of amateurs who are regularly making important research contributions.

Cosmic objects populate an enormous parameter space that spans many orders of magnitude in multiple dimensions, creating serious technological challenges for survey designers and instrument builders. At the top of the list of problems that must be addressed by any survey is the basic tension between the sky coverage, depth (sensitivity), and temporal resolution. The need for more wavelength coverage and better spectral resolution introduces another level of complication, hence a large diversity of approaches and designs. In the limited space of this chapter, we will not be able to even mention all relevant projects and contributions. Our goal is to illustrate the range of possibilities with examples chosen to demonstrate a particular point. The continuing pressure for more coverage, better completeness, and greater fidelity of the recorded information coupled with the availability of relatively cheap detector arrays is the driving force behind a data tsunami that is beginning to overwhelm the existing data processing capacity. The demand for data-intensive computing right at the telescope, hands-free data acquisition, and automated real-time knowledge extraction is greater than ever before. Taking humans out of the loop not only increases the efficiency of survey operations but is also an enabling factor for time-critical applications. Full automation is especially important for studies of explosive phenomena such as gamma-ray bursts (GRBs) that require rapid follow-up observations before the optical signal fades away on a time scale of minutes. Robotic telescopes and autonomous observatories completely dominate this arena. With the advent of the new generation of gravitational wave and neutrino observatories (e/a-LIGO, Icecube), we are witnessing the birth of multimessenger astronomy. Robotic instruments will play a key role in maximizing the scientific impact of these new detectors. Historically, observations outside the optical window provided the signatures of the most exotic physical phenomena, but the optical cross-identifications are still crucial for providing key ingredients of the physical interpretation such as accurate position and distance.

The golden era of great sky surveys is upon us. Over the past decade, we have seen rapid progress toward a continuous photometric record of the optical sky (Paczyński 2000). Numerous sky surveys are discovering and monitoring variable objects by hundreds of thousands, and innovative follow-up programs that were not possible just a few years ago are now in routine operation. Advances in detector, computing, database, and networking technology are enabling applications of all shapes and sizes ranging from small all-sky monitors, through networks of robotic telescopes of modest size, to big glass facilities equipped with gigapixel CCD mosaics. The Large Synoptic Survey Telescope (LSST) will be the first peta-scale astronomical survey (LSST Science Collaborations et al. 2009). It will expand the volume of the parameter space available to us by three orders of magnitude and explore the mutable heavens down to an unprecedented level of sensitivity.

LSST will deliver up to $\sim 10^5$ variability alerts per night. It will take a worldwide effort and a follow-up program coordinated on a global scale to utilize this amount of real-time actionable information. Automated scheduling and rapid response optimization will be the key to success. Telescope networking, the field that hardly existed 15 years ago, is taking survey astronomy by the storm. A brief review of publications on ADS reveals an explosion in the number of papers related to this topic. The words “telescope network” appear in nearly 3,000 abstracts with two thirds of the papers published in the last decade. And robots are doing the work. Roughly 1,300 abstracts contain the words “robotic telescope,” and over 80% of these were published after 2000. A good place to start for those who would like to learn more about robotic astronomy and telescope networking are the proceedings of the Heterogeneous Telescope Networks (HTN) workshop and Hot-Wiring the Transient Universe meeting published in special issues of *Astronomische Nachrichten* (Naylor et al. 2006; Allan et al. 2006a). Another valuable resource

on the topic is the special issue of *Advances in Astronomy* devoted to robotic astronomy (Castro-Tirado 2010).

2 Design Considerations for Autonomous and Survey Telescopes

This section summarizes the main factors that influence the design of robotic and survey telescopes given the required degree of automation. Ever since the introduction of computer-controlled electromechanical interfaces more than three decades ago, the level of automation in telescope operations and survey observing has been steadily increasing. It is hard to imagine a modern telescope without any robotic features. The spectrum of capabilities extends from simple command line interfaces that execute commands as they are typed by the observer all the way to fully autonomous systems requiring no human supervision to carry out their nightly activities. Today, the term robotic telescope is typically reserved to describe a complete observatory including the telescope with supporting instrumentation and control systems making observations without human intervention. In some cases, a human observer must initiate observations in the evening and closeout in the morning. By contrast, the so-called remote telescope, although operated by a human from a distant location using a computer network, does not qualify as a robotic telescope.

Robotic telescopes are complex hardware/software systems composed of a number of sub-systems such as optical elements, detectors, pointing devices, protective enclosures, weather monitoring equipment, and control logic. The main reasons to go through the trouble of building an autonomous observatory are:

- Streamlining survey operations to minimize gaps in observing and maximize efficiency
- Taking humans out of the loop to enable rapid response with very short reaction times
- Deploying instruments to remote sites with minimal infrastructure support or hostile environment, e.g., in order to achieve global sky coverage, while avoiding expensive travel

It is impossible to list all factors that may affect hundreds of decisions involved in this process, just as it is impossible to think of all potential applications for robotic telescopes. What follows is therefore an attempt to identify the most common scenarios.

2.1 Modular Design and Logistics

Early implementations of robotic telescopes built in the 1980s were very expensive when compared to their capabilities. The lack of reusable hardware and software contributed to very slow progress at that time. Over the past decade, the landscape has been completely transformed. The availability of inexpensive commercial off-the-shelf (COTS) components is fueling an explosion of activity in the area of robotic astronomy. While purpose-built components can be highly optimized toward specific research goals, the use of COTS equipment whenever available dramatically cuts the development time and cost. The ease of integration is paramount. Top level control systems are generally implemented in software that must communicate with all hardware components. When it comes to purchasing hardware, the availability of standard development kits and drivers for a particular operating system should always be considered.

Another important factor is open access to source code and specifications of signaling protocols for commercial products such as cameras, mounts, shutters, focusers, etc. Reliance on closed source software and proprietary standards is better avoided, as it frequently leads to a loss of flexibility to modify the design in the future and accommodate new observing programs. This is not always possible, however, and numerous robotic telescopes operating today utilize a mix of COTS and custom-made elements with some dependency on proprietary products.

Logistical considerations have a surprising way of creeping into every aspect of development in robotic astronomy. Availability of cranes at a remote site, the size of commercial shipping containers, and differences in voltage standards may not be the first things that come to mind when planning a new sky survey, but they have a great potential to become showstoppers if ignored. Shipping fewer preassembled modules is a way to eliminate complicated and labor-intensive jobs at the site. At the same time, this will increase the size and weight of parts that must be handled. If multiple telescopes must be installed around the globe, adding a voltage switch and replicating the same hardware is far easier than dealing with different components at different sites.

2.2 Optical Design

A vast majority of contemporary robotic telescopes are small instruments with some notable exceptions like the Liverpool Telescope (Steele et al. 2004). **Table 2-1** shows the distribution of apertures from a listing of 224 robotic telescopes¹ published by Frederic Hessman. The complexity of engineering an autonomous mechanical system increases very quickly with the size and weight of the structure, especially for applications that require fast slewing. However, our knowledge of the bright sky is still surprisingly incomplete (Paczyński 2001). Numerous science goals can be accomplished using commercially available photo lenses for very wide-field imaging (8–50 deg across). But there is a price for this convenience. Such systems deliver images dominated by instrumental seeing that may be significantly undersampled for the typical 14- μ CCD pixel. Another complication is strong vignetting with light attenuations reaching 50–60% at the edge of the field. The sky background is never uniform on such large spatial scales, so good flat field corrections are very difficult to obtain. Field distortions near image corners can reach the amplitude of 10–20 pixels. Standard data reduction and astrometric/photometric calibration techniques are usually of little help, and special algorithms must be developed on case by case basis. Introducing color filters and other specialized equipment such as polarizing and dispersing elements may be complicated by the fixed form factor. For some applications (e.g., localization of fast optical transients with broad emission spectra), skipping filters is a cheap way to boost the sensitivity of a small optical telescope, provided that the loss of fidelity and nonstandard calibrations can be tolerated. Stray light from the Moon is a serious problem for wide-field imagers and must be managed with a combination of filters and baffles.

At the next level of the size spectrum (0.2–0.5 m), there are several models of telescopes marketed primarily to amateurs and educators (Meade, Takahashi, DFM) that can be sufficiently ruggedized for professional use and autonomous operation (e.g., Vestrand et al. 2008; Bakos et al. 2009). But custom designs are also common in this aperture range. Larger telescopes are almost exclusively custom designed for particular projects. In wide-field imaging, the S/N ratio of the recorded point sources is dominated by the sky brightness due to a large angular size of

¹<http://www.uni-sw.gwdg.de/~hessman/MONET/links.html>

■ Table 2-1

Robotic telescopes of the world by aperture

Aperture (m)	# telescopes	Fraction (%)
<0.25	93	41.5
0.25–0.50	66	29.5
0.50–0.75	15	6.7
0.75–1.00	25	11.2
1.00–1.25	7	3.1
>1.25	18	8.0

■ Table 2-2

Data collecting power (etendue) of various telescopes

Instrument	Aperture (m)	FOV (deg)	$A\Omega$ (m ² deg ²)	N_{tel}
CONCAM	0.004	180.0	0.3	1
Super-WASP	0.100	15.0	11.1	8
ROTSE-III	0.453	2.0	0.5	1
RAPTOR-P	0.143	15.0	11.3	4
RAPTOR-T	0.420	0.4	0.02	1
RAPTOR-Q	0.017	60.0	3.1	5
ASAS	0.100	3.0	0.06	1
HAT	0.100	8.0	0.4	1
LINEAR	1.000	1.4	1.2	1
SDSS	2.500	1.5	8.7	1
Palomar/QUEST	1.200	4.0	14.2	1
PS-1	1.800	2.6	13.5	1
LSST	8.400	3.0	391.7	1
KAIT	0.800	0.1	0.004	1

pixels that collect lots of sky photons in a short time. Narrow field follow-up instruments can afford larger plate scales and therefore longer exposures for the same aperture size. Factors such as these in combination with primary science requirements and budget constraints ultimately determine the type and number of telescopes to build. Depending on derived survey parameters such as sky coverage, frequency of observations, and magnitude limits, an array of smaller telescopes may be a better fit than a single instrument with a large detector.

Covering a large portion of the sky with sensitive observations in a short amount of time is hard. Survey grasp measures the rate at which a given instrument collects information about objects of interest, and is one of the most important characteristics of a survey telescope. A figure of merit commonly used to compare survey grasp of various telescopes is the etendue defined as $A\Omega$, the product of the collecting area times the solid angle covered by the instantaneous field of view. The larger the etendue, the more objects will be recorded in a single exposure, and less time will be taken to survey a given area of the sky. ● Table 2-2 compares parameters of several telescopes and telescope arrays: aperture diameter in meters, field of view in degrees,

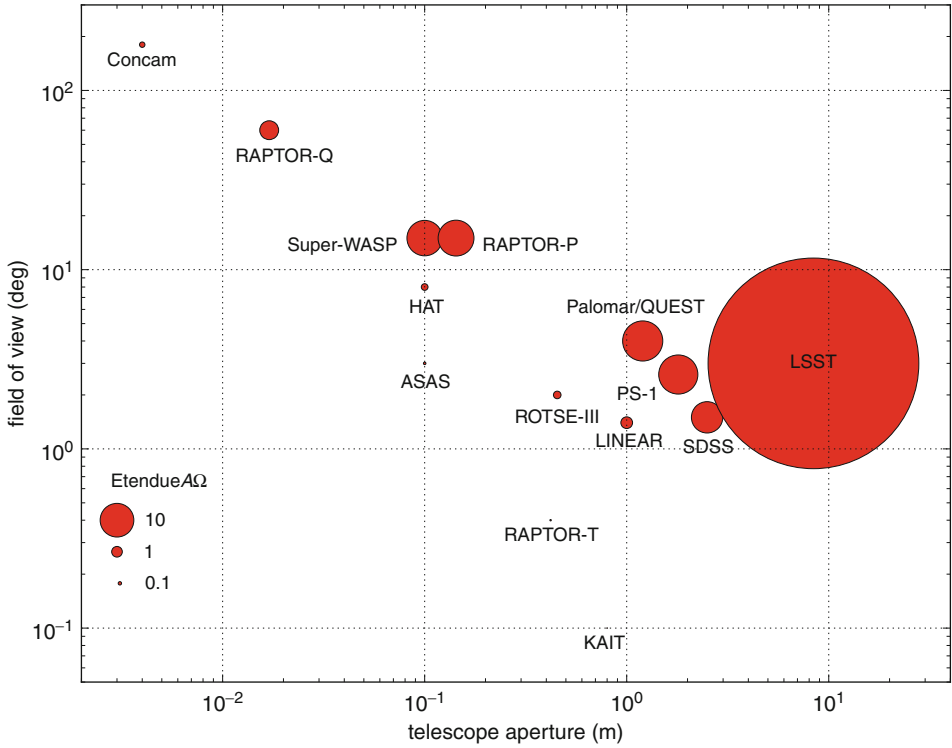


Fig. 2-1 Survey grasp comparison for selected robotic and survey telescopes from [Table 2-2](#). The chosen figure of merit is etendue equal to the product of the collecting area times the solid angle covered by the combined field of view for each instrument. The values range from 0.004 for KAIT to 390 for LSST and are proportional to the area of the corresponding symbols. Note that the etendue of small wide-field telescopes can be comparable to that of much larger survey instruments, albeit at a different sensitivity

etendue in meter degrees squared, and the number of individual telescopes N_{tel} . This data is also shown in [Fig. 2-1](#), where the area of the symbol is proportional to $A\Omega$. As expected, larger telescopes tend to cover a smaller field of view. More interesting is the fact that for certain types of surveys, relatively small telescopes and photo lenses occasionally outcompete much larger instruments. For example, four 200-mm Canon EF lenses have a larger data-gathering power than the 2.5-m SDSS telescope. Obviously, etendue does not tell a complete story. The former system offers great sky coverage at the expense of sensitivity, while the opposite is true for the latter telescope. Other measures of survey grasp may be more relevant depending on the science goals. In case of time-domain surveys, it is informative to compare the depth of the survey at a specified cadence, which gives an idea of the volume of the universe sampled for a particular type of transient variability. The number and type of filters to include with the telescope are also crucial for the science mission but are more difficult to include in survey metrics.

2.3 Telescope Mounts and Fast Slewing

Robotic telescopes are frequently deployed to perform rapid follow-up observations of targets at random locations and times. This drives the need for fast-slewing speeds. Even in the survey mode, faster repointing of the telescope will eliminate breaks between exposures and improve efficiency. In a sense, the fastest telescope is the one that does not have to move at all, because it already contains any possible target in its field of view. This is how full-sky monitors work. Such instruments are implemented using a set of stationary wide-field lenses or a single fish-eye lens for a full horizon to horizon imaging (➤ Sect. 3.1) and are quite limited in sensitivity. For larger pointed instruments, fast slewing at rates exceeding 10 deg/s poses the challenge of constructing a mount that is both rigid and not too heavy and delivers a high torque. ➤ Table 2-3 and ➤ Fig. 2-2 compare the peak slewing rate for several telescopes spanning a large range of apertures. Telescope flexure and oscillations in the acceleration/deceleration phase strongly limit the range of practical solutions for large telescopes. Custom-built fast-slewing mounts for military applications tend to be very expensive. ROTSE-III and several types of RAPTOR telescopes (➤ Sects. 3.2.4 and ➤ 3.2.5) are among the fastest slewing astronomical instruments ever built, achieving peak rates around 50 deg/s and accelerations reaching 20 deg/s² at ~0.4-m aperture. Both are based on a very successful line of mounts from The Pilot Group, LLC.

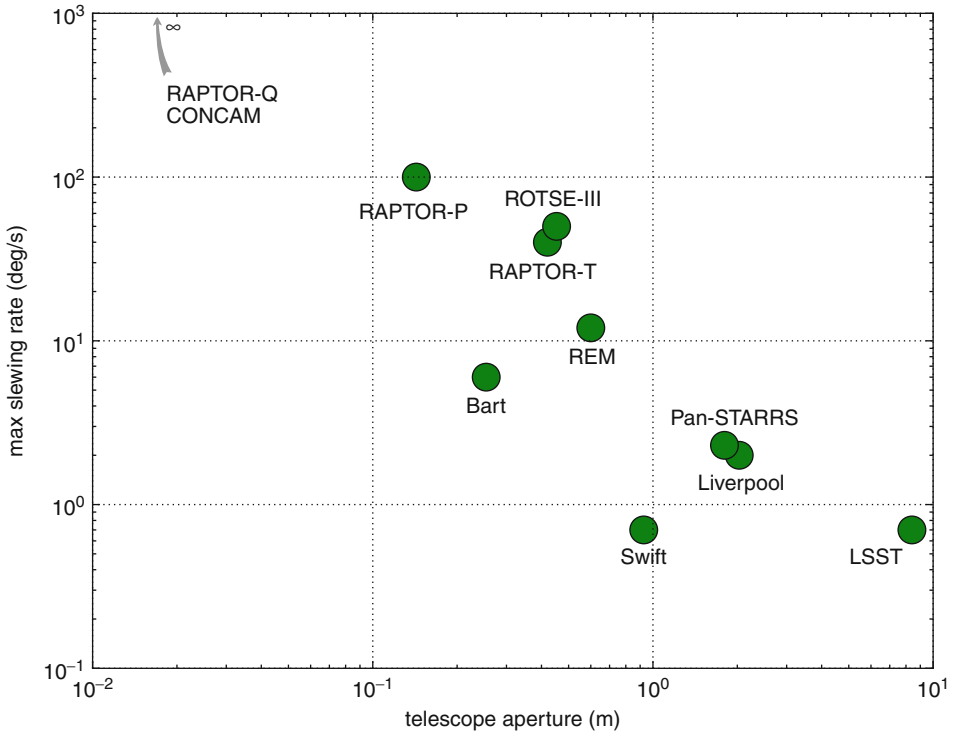
2.4 Detectors

The availability of affordable CCD cameras is arguably the single most important enabling factor behind the proliferation of robotic telescopes and autonomous observatories. Small format 512×512 pixel entry level science grade detectors prepackaged with readout electronics and thermoelectric cooling can be purchased for less than 10,000 dollars as of 2011. The Apogee Imaging Systems² product line called Alta U series consists of over 30 cameras including front- and

■ Table 2-3
Telescopes compared by slewing speed and size

Instrument	Aperture (m)	Slew rate (deg/s)
Bart	0.254	6.0
Liverpool	2.032	2.0
LSST	8.400	0.7
Pan-STARRS	1.800	2.3
RAPTOR-P	0.143	100.0
RAPTOR-T	0.420	40.0
RAPTOR-Q	0.017	∞
CONCAM	0.004	∞
REM	0.6	12.0
ROTSE-III	0.453	50.0
Swift	0.927	0.7

²<http://www.ccd.com>



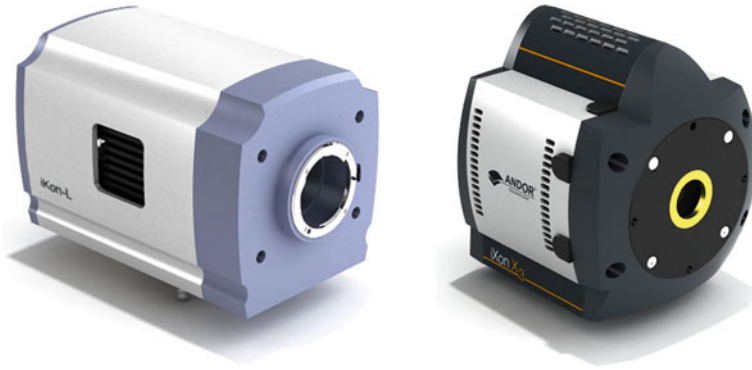
■ Fig. 2-2

Comparison of the peak slewing rate for robotic and survey telescopes of various apertures. All-sky monitors such as RAPTOR-Q and CONCAM are formally infinitely fast. The Swift satellite offers state-of-the-art slewing capability in space. The aperture for Swift is the diameter of the circular aperture equivalent to the collecting area of the onboard gamma-ray Burst Alert Telescope (BAT)

back-illuminated CCDs, as well as a few interline transfer models. Andor Technology³ offers high-quality cameras based on E2V chips including high-sensitivity electron multiplication CCDs (EMCCD) allowing extremely fast frame readout at video rates. This enables observations of astronomical sources with time resolution of 0.03 s or better or even in the photon counting regime in case of faint objects. Back-illuminated CCDs are characterized by much higher quantum efficiency and better response to blue light by comparison to front-illuminated chips. But they are about four times more expensive and may not offer any advantage for wide-field observations dominated by the sky background. ➤ Figs. 2-3 and ➤ 2-4 show examples of devices that found use in robotic telescopes. Any of these products can be connected to a computer over the standard USB 2.0 interface.

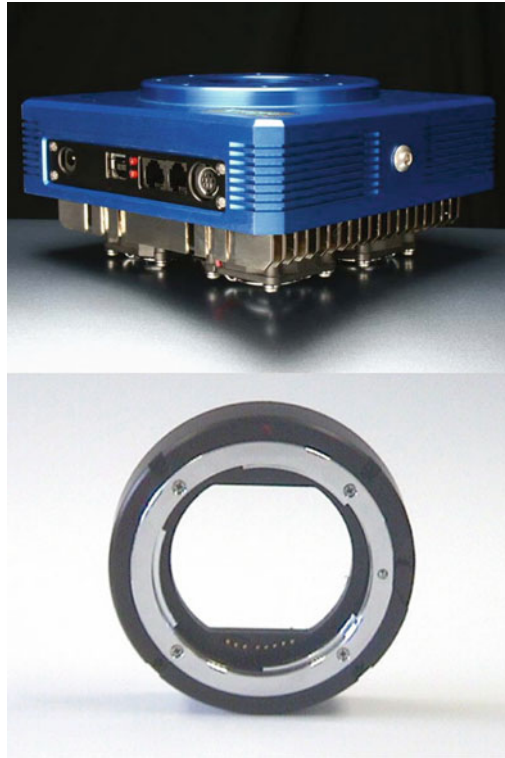
One of the most important elements of a CCD camera is the cooling system. Filling a dewar with liquid nitrogen (LN₂) on a daily basis is simply out of the question for autonomous systems. Air-cooled Peltier devices are very popular in small CCD cameras and used extensively in robotic astronomy. Closed-loop liquid cooling may be a better option for a larger detector.

³<http://www.andor.com>



■ Fig. 2-3

Prepackaged CCD detectors from Andor Technology: the iKon-L camera with a $2,048 \times 2,048$ pixel high dynamic range CCD (*left*) and model iKon X3 hosting an electron multiplication CCD (*right*)



■ Fig. 2-4

Alta U10 CCD camera from Apogee Imaging Systems (*top*) and the autofocus ring for Canon lenses from Birger Engineering Inc.⁴ (*bottom*)

⁴<http://www.birger.com>

Andor cameras for low light imaging applications now feature a five-stage thermoelectric cooler that can maintain temperatures as low as -100°C comparable to solutions based on cryogenics. Large surveys optimized for specific science objectives require purpose-built focal plane arrays that often employ dozens of CCD detectors (● Sect. 3.5).

2.5 Weather Protection

An autonomous robotic observatory must be able to protect itself from adverse conditions such as rain and high winds. Otherwise, sooner or later it will suffer a catastrophic damage. Therefore, most robotic telescope enclosures are connected to a dedicated weather station that monitors ambient temperature, humidity, barometric pressure, wind velocity, and, most importantly, precipitation. IR cloud sensors and sky brightness monitors are also frequently employed. Upon detecting conditions outside the operating range, the system automatically shuts down.

Another level of protection is required to hedge against the loss of power and computer/software failures. This can be accomplished with a separate controller, a watchdog timer that initiates an immediate shutdown unless it receives “I’m alive” signal every second or so. Emergency uninterruptible power supplies must be installed with sufficient capacity to close the roof and to provide clean power at remote sites. Many survey telescopes are housed in rotating domes that shield against the wind. However, this option does not work for rapid follow-up telescopes that must be able to point in a random direction on the spot. Several types of enclosures providing an unobscured view of the entire sky are shown in ● Fig. 2-5. A classic roll-off roof is one possibility. There are also many types of “clamshells” with one or more sections that rotate out of the way or fold to the side.

2.6 Control Systems, Data Acquisition, and Processing

Robotic and survey telescopes require a large amount of sophisticated control software to operate. Reliable operation without human supervision implies a high level of system self-awareness; multiple layers of automatic control, recovery, and parallel execution; robust hierarchical status reporting; and hardware redundancy (Bowman et al. 2002). To achieve the required level of reliability, such systems must monitor their own health. Ideally, faults would be predicted to give some advance warning when an equivalent of the “check engine” light in a car turns on. Structural health monitoring methodology has been applied to successfully predict mechanical failures in RAPTOR telescope mounts using simple sensors that “listen” and detect anomalous vibration patterns (Stull et al. 2011).

Large instruments typically require custom-built control systems. However, there are several software packages that have reached the level of a complete observatory manager and are being adopted by a growing number of autonomous observatories. The top-level architecture of such system is typically based on a set of daemons, each tasked with controlling and monitoring some hardware component: telescope, mount, camera, enclosure, and weather station. There are also daemons for scheduling using a variety of inputs such as precompiled list of targets, live alert feeds from external sources, or even manual override. The central master process communicates with the rest of the system via predefined protocols (e.g., string commands over TCP sockets or shared memory, XML-RPC, HTTP). RTS2, the second-generation Remote



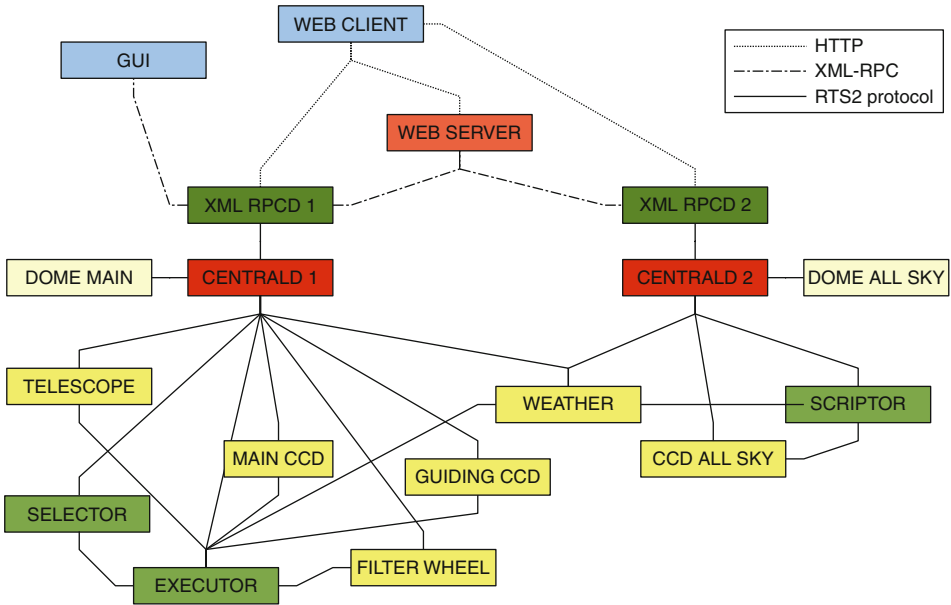
■ Fig. 2-5

Comparison of enclosure designs for autonomous observatories and survey telescopes. (a) A symmetric clamshell of ASAS actuated by a commercial gate motor. The rotating dome of the 1.3-m Warsaw telescope with adjustable slit is visible in the background. (b) Highly portable ROTSE-III enclosures with a one-sided clamshell consist of four easy to assemble pieces. (c) This node of the MASTER telescope network is an example of a popular dome design that folds on itself when opening. (d) Classical roll-off roof enclosure used by SuperWASP

Telescope System,⁵ is an open-source package for managing networks of autonomous robotic observatories with the goal of providing plug and play functionality (Kubanek 2010b). A block diagram of a particular system under the control of RTS2 is shown in ● Fig. 2-6. This software has been deployed on over dozen telescopes and runs on a wide variety of Linux distributions, the Solaris operating system, and at least in part on Windows and Mac OS X.

Optimal scheduling of observations is a complex problem. The list of commonly used algorithms includes (1) precompiled night plan; (2) dispatch scheduling, i.e., maximizing the merit function for the next target; and (3) semi-automatic queue scheduling of observable targets with manually adjusted priority. In addition to those three methods, RTS2 offers scheduling based on constrained optimization over a set of possible schedules using genetic algorithms (Kubanek 2010a). Most robotic systems have some provisions for observing targets of opportunity (TOO). The simplest response protocol is to abort the current observation, immediately slew to a new

⁵<http://rts2.org>



■ Fig. 2-6

Schematic view of a hypothetical system controlled via RTS2 (second-generation Remote Telescope System). RTS is an integrated open-source package for automated observatory control

target and execute a preprogrammed sequence of exposures. Autonomous observing is rapidly evolving to include increasingly sophisticated data processing tasks such as astrometric and photometric pipelines, automated detection of variability, and real-time cross-identification of sources. One of the most important developments of the last decade is closed-loop observing whereby the results of the automated data analysis, system status, and quality assurance information are fed back into the system and used to schedule a tailored follow-up sequence. In 2002, a pair of wide-field telescopes deployed by the RAPTOR project at Los Alamos National Laboratory became the first robotic system operating in a closed loop. Coordinated observations of rapidly variable objects (e.g., GRB follow-up) require a time reference accurate to 0.1 s or even better. This is typically achieved with a combination of NTP (Network Time Protocol) servers and GPS receivers. Network bandwidth is a scarce resource at a typical observing site and forces time-sensitive and data-intensive projects to process the bulk of the data right at the telescope. Many robotic observatories are now equipped with substantial computing and storage resources.

3 Diversity of Survey and Follow-Up Instruments

3.1 All-Sky Monitors

All-sky monitors populate the bottom of the telescope food chain. They are small and limited in sensitivity but also cheap and easy to replicate for what they can deliver. The idea is to have

a complete “passive observatory” designed to work without human supervision that provides a truly continuous record of the night sky. Instruments in this category see from horizon to horizon but do not track the sky. Although they are mostly constructed out of COTS components, significant modifications are typically required to put together a complete system.

3.1.1 CONCAM

The Night Sky Live project (Pérez-Ramírez et al. 2004) has deployed three generations of CONCAM (continuous camera; ● Fig. 2-7). The system consists of a fish-eye lens attached to a CCD camera under the control of a laptop computer. The main design principle adopted by the creators of CONCAM is “If it moves, it breaks,” so there are no moving elements in this design, except for the rotating hard drive of the control computer. A complete remotely controlled observatory fits in a container the size of a large briefcase and is designed to be bolted to a rooftop. Depending on the system version, the optical element is either Nikon FC-E8 or SIGMA F4-EX 8-mm fish-eye lens with FOV ~ 180 deg. A low-cost CCD camera with a relatively low dark current was chosen to allow long integrations: SBIG ST-8 (CONCAM-2) or ST-1001E (CONCAM-3). Any low-end personal laptop has sufficient computing power, storage, and network connectivity to serve the needs of CONCAM. All components are mounted inside a modified case from Pelican to make a self-contained package that is easy to transport, maintain, and operate. Power, Ethernet, and telephone wires connect to the bottom and are threaded through the center of a mounting tube.

Every 236 s, CONCAM takes a single $1K \times 1K$ pixel image exposed for 180 s that contains approximately 1,200 objects down to a detection limit slightly better than that of the human eye. CONCAMs were the first source of real-time all-sky relative opacity maps and resolved cloud cover information to support astronomical observing sites, including Gemini North, Keck, Subaru, IRTF, Spacewatch, Wise, ING 4-m, Mayall 4-M, SARA, and WIYN. Photometric and astrometric calibration of this type of data poses a challenging problem and typically requires




■ Fig. 2-7


CONCAM installation at the Siding Spring Observatory near Coonabarabran, Australia

custom algorithms and software. Of some help is the fact that for an instrument pointed at zenith, many lens and sky characteristics become functions of the zenith angle alone. CON-CAM has been used to derive limits on bright optical counterparts to GRBs. It is well suited for a variety of science programs including observations of meteors, variable stars, novae, and supernovae, as well as education and outreach activities.

3.1.2 RAPTOR-Q

The RAPTOR/Thinking Telescopes network developed at the Los Alamos National Laboratory hosts telescope arrays of several types covering a range of capabilities (Vestrand et al. 2008). RQD2 (Raptor-Q Design 2) is a small, transportable, robotic observatory designed to function as the basic node in a global network capable of continuous persistent monitoring of the night sky (Wren et al. 2010). A single unit ready for deployment is shown in  Fig. 2-8. The observatory employs five wide-field imagers that altogether view about 90% of the sky above



 Fig. 2-8

RQD2, an autonomous all-sky monitor developed at the Los Alamos National Laboratory. The enclosure contains a complete observatory with control computers, data processing, and storage hardware in a compact air-conditioned package

12 deg elevation with a sensitivity of $R \sim 10$ mag assuming a 10-s exposure. This system fills an important niche in the field of robotic astronomy, providing nearly all-sky monitoring with the sensitivity and cadence suitable to detect interesting astronomical transients lasting as little as 30 s. Another important use of RQD2 is as an extinction and cloud cover monitor with sufficient spatial resolution to allow other facilities navigate in partially cloudy conditions.

The most expensive part of the system is a set of five CCD detectors. Thermoelectrically cooled Alta U10 cameras from Apogee featuring a $2,048 \times 2,048$ pixel front-illuminated CCD provide good image quality at a reasonable cost. The pixel readout is digitized at 16 bit resolution, and an entire image can be read out in 4 s. The camera is controlled over a USB 2.0 connection. Although Apogee does not provide LINUX drivers for its products, the Random Factory based in Tucson AZ provides a LINUX driver for these cameras. The sky is imaged using off-the-shelf Canon 24-mm EF camera lenses with a focal ratio of 1.4 and a 17-mm clear aperture. This particular camera/lens combination covers a 60-deg field of view. All cameras are attached to the enclosure and pointed in fixed directions, four of them at a 45-deg elevation above the horizon, and the fifth at zenith. Canon EF lenses allow autofocus using an ultrasonic motor built into the lens. Birger Engineering Inc. has reverse engineered the proprietary lens control system and offers a device for controlling the Canon lenses via a serial link. In RQD2, the Birger units connected to the camera control computer are used to compensate for focus changes due to variations in the ambient temperature.

The RQD2 enclosure must provide complete environmental protection for lenses, cameras, and substantial computing resources. It is constructed out of two combined 4/12 NEMA (National Electrical Manufacturers Association) enclosures, one $24 \times 24 \times 16$ and the other $24 \times 24 \times 8$ in. in size. A major design goal was to make the system easily transportable and rugged enough to survive shipping to distant locations. The NEMA enclosures are standard off-the-shelf components constructed of 14 ga steel and have a watertight closed-cell polyurethane seal on the doors. The “dome” is a custom designed truncated pyramid structure made of 14 ga aluminum sheet metal and attached to the top of the NEMA box cube. CCD cameras with autofocus mechanisms are attached to the inside faces of the dome. The lenses are protected with cover boxes, small $8 \times 6 \times 5$ in. NEMA 4 enclosures fitted with circular aluminum shutters. The shutters actuated by 12 VDC gear motors are the only moving parts in this system (except for hard disk platters in control computers). RCM-4000 microcontroller from Rabbit Semiconductor Inc. is used to command the five individual cover controllers (PIC 16F88) over the RS-485 serial link. During observations, the shutters are rotated away from the optical axis to provide unobstructed field of view. The lens compartment is insulated with closed-cell silicone foam. A watchdog timer closes the covers within 10 s in case of a system lockup or a software error. The system is programmed to automatically stop observations and cover the lenses in bad weather conditions. The weather station on RQD2 includes a precipitation detector (Vaisala DRD11A), an anemometer, and a temperature/humidity sensor monitored by a separate Rabbit RCM4000 controller running a TCP/IP server over its Ethernet interface. The system is powered using a 1,000-W 12 VDC power supply capable of running on 120/240 VAC, 60/50 Hz, grids and backed up by an uninterruptible power supply (UPS). An external 10,000 BTU air conditioner is normally connected to the main RQD2 enclosure via 5” diameter conduit to keep the interior temperature in the operating range.

RQD2 performs full online data processing using its onboard network of six computers running Linux. Each custom-built computer consists of a Mini-ITX motherboard, an Intel 2.5-GHz Core2 Duo processor, 4 GB of RAM, and a 250-GB hard drive installed in a standard 1U rack mount. The master node controls five slaves (one per camera) via the USB interface. This way,

the cameras can be exposed and read out in parallel, and the data analysis can be completed before the next exposure is initiated. The typical exposure time is 10 s, followed by a 4–5 s read-out. Images are taken 20 s apart and stored in a 3TB RAID array. Computers are cross-mounted using the Network File System (NFS) and synchronized with the Network Time Protocol (NTP). A Virtual Network Computing (VNC) server provides remote access to machines that may require a soft reboot, while the power distribution unit (PDU) allows computers to be remotely power cycled.

The first-generation RAPTOR-Q system consists of cameras mounted to enclosures of larger RAPTOR telescopes and has been in operation since 2006. Despite its lower image quality due to the lack of autofocus mechanism, and therefore lower sensitivity, it obtained scientifically important observations such as the onset of the eruption on comet 17P/Holmes that brightened by a factor of 10^6 , thus demonstrating the potential of RQD2. Operating autonomously, RQD2 acquires a nearly full-sky image every 20 s, taking more than 10,000 individual $2K \times 2K$ pixel frames per night. It also runs real-time astrometric and photometric pipelines that provide the capability to search for bright astronomical transients and monitor transparency changes across the full sky. The first RQD2 observatory was deployed in March 2009 and is currently operating at the Fenton Hill site near Los Alamos, New Mexico.

3.2 Mid Range Robotic Telescopes

3.2.1 ASAS

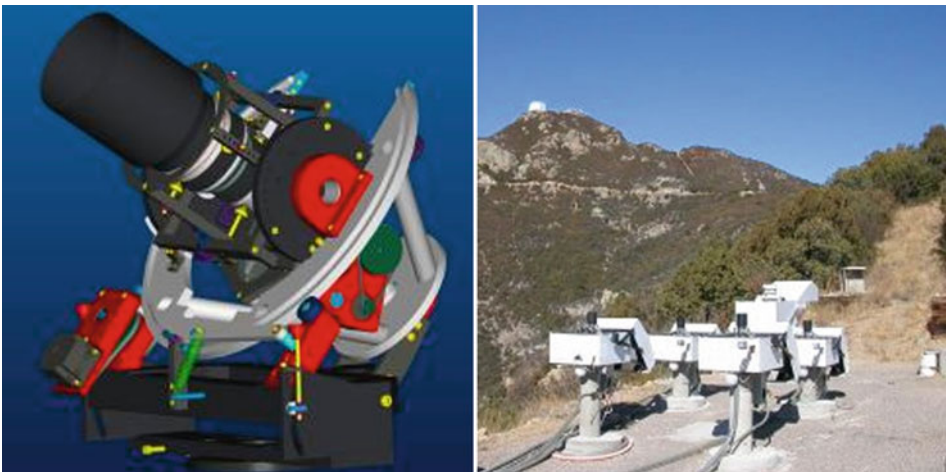
One of the pioneering efforts in the area of continuous photometric sky monitoring for variability is the All-Sky Automated Survey (ASAS) (Pojmanski 1997; Pojmański 2009). The project began in 1996 and was in part inspired by Prof. Bohdan Paczyński. The idea was to use inexpensive COTS components to build several small telescopes capable of surveying the bright sky in a completely automated manner. The first ASAS device at Las Campanas, Chile, was a 135-mm $f/1.8$ photo lens with a 768×512 pixel amateur grade CCD camera on a horseshoe parallax mount that could point in about 10 s and track for 5 min. In 2000, this was replaced by two wide-field cameras equipped with 200-mm $f/2.8$ lenses and $2K \times 2K$ pixel CCDs covering $8.5 \times 8.5 \text{ deg}^2$ each and a 250-mm $f/3.3$ modified Cassegrain reflector for follow-up. ASAS-North wide-field instruments in Hawaii attached to the platform of the Faulkes Telescope North are similar, except for better Nikon 200 $f/2.0$ lenses. Early on, the system required some assistance from observers at the nearby OGLE telescope, but after a move to computer controlled domes, human support has been reduced to occasional maintenance and data transport. Fields are selected from a fixed list by automated scheduling software based on their position with respect to zenith, Moon, and desired frequency of observations. Each camera takes between 160 and 250 images per night, and the entire available sky is covered in two filters after 1–2 days. ASAS quickly demonstrated that at least 80% of bright variable stars are still waiting to be discovered. In the first decade of operation, ASAS collected 700,000 images in V, R, I bands and constructed a photometric catalog of 20 million stars brighter than $V = 14.5$ with hundreds of measurements per star on average. The resulting catalog of 50,000 variable stars, mostly brighter than 12.5 mag, is the first homogeneous record of this kind. The ASAS Alert Service provides almost instantaneous notification on unexpected behavior of existing or new objects above 12.5 mag limit.

3.2.2 HAT

HAT (Hungarian Automated Telescope) started as an all-sky variability search of the northern hemisphere (Bakos et al. 2002). HATNet established in 2003 gradually evolved into a multi-site, multi-instrument network with instruments deployed at FLWO, Mauna Kea, and also WISE Observatory, searching for transit signatures of extrasolar planets as they “eclipse” their parent star. The first-generation network was composed of stand-alone, fully robotic units, each consisting of a 200-mm Canon telephoto lens with an *I*-band filter and a $2K \times 2K$ pixel front-illuminated Apogee AP10 CCD on a horseshoe mount (📍 Fig. 2-9). The latest generation HAT-South experiment is designed to bridge the gap between shallow and medium-deep searches for transits of planets around F0-M5 dwarfs (Bakos et al. 2009). With three sites in Chile, Australia, and Southern Africa, the network provides nearly round the clock monitoring of selected fields, which is critical for achieving high efficiency to detecting brief light curve dips that last a few hours and repeat every few days. Each site hosts two TH₄ arrays of four 0.18-m *f*/2.8 Takahashi hyperbolic astrographs with Apogee Alta U16 4K \times 4K CCDs. The effective field of view of the array is 8×8 deg and the pixel scale is 3 arcsec/pixel, which greatly limits the impact of source confusion that dilutes the eclipses and increases the number of false positives to be rejected with follow-up observations. The telescopes are protected by a clamshell dome and operated in a fully automated manner. Since 2006, HAT discovered 30 planets. HAT-South is expected to generate roughly 250 candidate transiting planets per year.

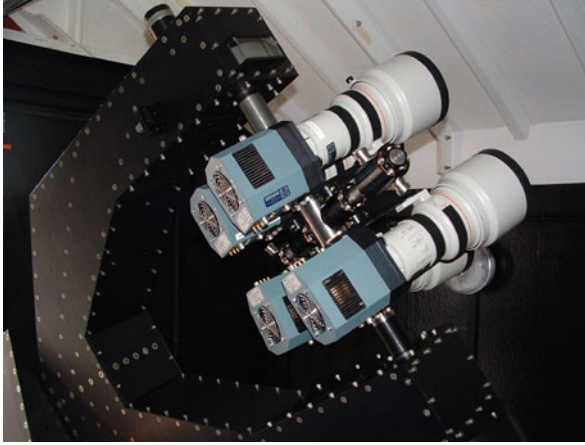
3.2.3 WASP and SuperWASP

The Wide Angle Search for Planets (WASP) is one of the leading exoplanet transit detection programs conducted by a consortium of eight academic institutions in UK (Pollacco et al. 2006;



📍 Fig. 2-9

Hungarian Automated Telescope (HAT): design of a single unit (*left*) and an array of four units deployed at the FLWO, Arizona (*right*)



■ Fig. 2-10

Array of wide-field telescopes and CCD cameras employed by the WASP project

Christian et al. 2006). The SuperWASP instruments are located on La Palma, Canary Islands, and at the South African Astronomical Observatory. Each station is equipped with an array of eight wide-field cameras on a single equatorial mount capable of slewing at a rate of 10 deg/s inside a roll-off roof enclosure (▶ Fig. 2-10). The CCD cameras based on back-illuminated $2K \times 2K$ pixel E2V chips are from Andor Technology. Like several other projects of this type, SuperWASP uses Canon 200-mm f/1.8 telephoto lenses and a broadband visual filter (400–700 nm) for high-quality wide-field imaging. Each array covers a combined field of view of 482 deg^2 sampled at $\sim 14 \text{ arcsec/pixel}$ and is capable of surveying the entire visible sky every 40 min. The combined data rate is up to 100 GB per night, depending on the observing strategy. The top level control software is a modified version of the commercial Linux package Talon from Optical Mechanics Inc. that has been released in the open-source domain. Observing is fully automated and the telescopes can protect themselves in case of a weather alarm. Failure to close the roof generates a radio signal for an attendant in a neighboring dome. The system routinely delivers photometry accurate to 1% for objects in the magnitude range $V \sim 7.0\text{--}11.5 \text{ mag}$ and so far discovered over 30 planets.

3.2.4 ROTSE

Robotic Optical Transient Search Experiment (ROTSE) dates back to the mid-1990s when the first-generation instrument was installed at the Los Alamos National Laboratory, New Mexico. The system consisted of four Canon lenses attached to commodity large-format CCD cameras on a fast-slewing mount and covered the field of view $16 \times 16 \text{ deg}$ to match GRB localization error boxes delivered by the BATSE satellite. On January 23, 1999, ROTSE-I made a landmark observation, the first detection of contemporaneous optical emission from an exceptionally bright GRB that reached ninth magnitude in visible light (Akerlof et al. 1999). This result greatly stimulated further development of robotic GRB follow-up projects. The current generation ROTSE-III system is a global network of 0.45-m telescopes operated by an international consortium (Akerlof et al. 2003). With telescopes located in Australia, Namibia, Turkey, and Texas,



■ Fig. 2-11

ROTSE-IIIb telescope on Mt. Fowlkes, Texas. ROTSE-III is a network of four virtually identical 0.45-m telescopes on fast-slewing mounts capable of observing poorly localized transients within a short few seconds following the alert

the ROTSE-III network provides 24-h coverage of the night sky. The telescope is a modified Cassegrain with a very fast $f/1.8$ primary mirror and a refractive field corrector on a rapidly slewing mount (► Fig. 2-11). Achieving a large field of view (2.64 deg diameter) was an express design goal driven by the accuracy of GRB positions from HETE. The $2K \times 2K$ pixel back-illuminated Marconi CCD detector with $13.5\text{-}\mu$ pixels in combination with the effective focal length of 0.85 m delivers pixel scale of 3.28 arcsec/pixel and produces undersampled images of point sources. This was accepted in order to meet the specification on the field coverage. The ROTSE-III camera from Astronomical Research Cameras employs propylene glycol heat transfer loop driven by air-cooled recirculator to cool the CCD sensor down to -40°C from ambient temperatures reaching 20°C . The CCD detector and the field corrector are mounted inside the telescope tube between the primary mirror and the flat secondary. The telescope mount can slew at a maximum speed of 35 deg/s and accelerate at 16.4 deg/s^2 in right ascension and 20.6 deg/s^2 in declination. This makes it possible to slew from horizon to horizon in just 8 s. A typical response time to a GRB localization is less than 4 s from a standby position (zenith). In order to collect as many photons as possible, ROTSE-III telescopes are operated without filters. Since the launch of the Swift satellite in the fall of 2004, ROTSE telescopes are a leading supplier of early detections and accurate positions of optical GRB counterparts. They were also used in a successful supernova search.

3.2.5 RAPTOR/Thinking Telescopes

Since the very beginning in 2001, the RAPid Telescopes for Optical Response (RAPTOR) project at Los Alamos National Laboratory has been focusing on innovative approaches to telescope networking and untriggered searches for optical transients. The first-generation system was a pair of identical arrays of photo lenses with CCD cameras. It was designed to mimick the functionality of evolved biological systems such as human vision (Vestrand et al. 2004b).

Each array had four wide-field detectors (corresponding to rods of a human eye) and a single more sensitive, narrow-field detector in the middle (fovea) for immediate self follow-up after a slight change of pointing (closed-loop operation). RAPTOR-A and B were separated by 38 km to enable selection of transient objects in the “stereoscopic vision” mode. This allowed the system to reject artifacts and false positives due to man-made phenomena based on parallax measurements and coincidence filtering. Real-time, automated data analysis pipeline served as the “brain.” In 2002, RAPTOR became the first fully autonomous robotic telescope to operate in a closed loop.

The Thinking Telescopes project is a continuation of the original RAPTOR program with the purpose of developing a global network of sky monitors and rapid response telescopes to search for and follow up optical transients in real time (Vestrand et al. 2008). This is a challenging goal that will require a sustained investment in information technology and software (Vestrand et al. 2004a). The current system includes several types of telescopes optimized for different roles (● Fig. 2-12). Fast-cadence/full-sky monitoring is accomplished with RAPTOR-K and RAPTOR-Q. RAPTOR-K is a wide-field array for persistent optical monitoring composed of 200-mm Canon lenses carried on a single rapidly slewing mount. Each telescope has a $2K \times 2K$ pixel E2V back-illuminated CCD at the focal plane and detects sources down to a 3-sigma limiting magnitude of $R \sim 16$ mag in 30 s. Altogether, the full array of 16 telescopes covers an instantaneous field of view of $1,024 \text{ deg}^2$ and is capable of patrolling the full sky on a 15-min circuit. RAPTOR-Q persistent monitoring arrays (● Sect. 3.1.2) provide a continuous record of the night sky to a depth of ~ 10 mag every 20 s. RAPTOR-T (for technicolor) is tasked with fast response and multicolor imaging. These follow-up arrays are composed of four 0.4-m $f/5$ optical telescopes that are co-aligned and colocated on a single rapidly slewing mount. Each of the four channels images through a different filter (Johnson *VRI* and clear). An advantage of this approach over a traditional filter wheel is that it allows true simultaneous multicolor



■ Fig. 2-12

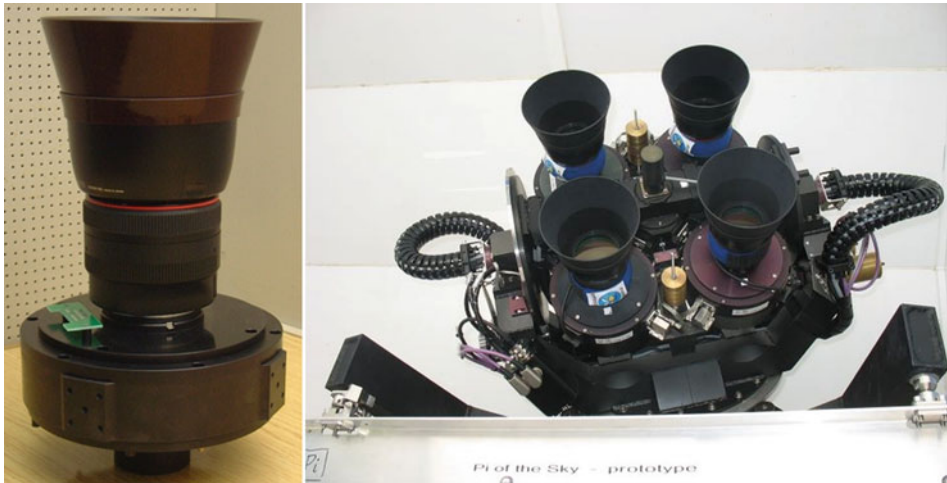
RAPTOR/Thinking Telescopes arrays at the Fenton Hill observatory, New Mexico: RAPTOR-K covering $1,000 \text{ deg}^2$ down to ~ 16 mag designed to autonomously detect optical flashes on time scales of minutes (*left*) and RAPTOR-T featuring four co-aligned 0.4-m telescopes for simultaneous follow-up imaging in *VRI* and clear channels (*right*)

imaging of rapidly varying sources. RAPTOR-T arrays slew to the location of the transient and begin observations in less than 10 s. They respond to both internal triggers from RAPTOR-K/RAPTOR-Q and external alerts such as GRB localizations from GCN (GRB Coordinates Network; cf. [Sect. 4.1](#)). This unique combination of simultaneous multicolor imaging, relatively large aperture, and short response time makes RAPTOR-T the instrument of choice for probing the prompt optical emission of GRBs within a few seconds of the explosion.

3.2.6 “Pi of the Sky”

The “Pi of the Sky” (PIOTS) experiment is designed for continuous high-cadence observations of a large region of the sky and searching for astrophysical variability on short time scales down to a few seconds (Malek et al. 2010). The project builds their own custom designed CCD cameras employing ST A0820 2K × 2K pixel chips cooled by a two-stage Peltier module. Another interesting innovation introduced by PIOTS is a highly durable shutter based on the voice-coil principle used in read-write heads of hard disk drives. By eliminating friction, the durability was dramatically increased to about 10^7 opening cycles (over 2,000 exposures per night for a few years). This additional investment of time and effort resulted in both better performance and substantial savings in hardware cost. PIOTS uses 85-mm, f/1.2 Canon EF lenses to detect point sources down to ~ 12 mag in a single unfiltered 10-s exposure. The frames are separated by readout intervals lasting only 2 s.

Transient detection runs in coincidence mode between a pair of detectors covering the same 20×20 deg field of view. [Figure 2-13](#) shows the camera and a single PIOTS array. The prototype system operating since 2004 at the Las Campanas Observatory, Chile, has been very



■ Fig. 2-13

“Pi of the Sky” detector system: optical element attached to the camera (*left*) and the first array installed at the INTA El Arenosillo test center in Mazagon near Huelva, Spain. The main goal of the “Pi of the Sky” experiment is to detect optical counterparts of GRBs before and during the gamma emission

successful. Its most important contribution is the discovery of the optical transient associated with GRB 080319B that peaked at 5.3 mag, the “naked-eye burst,” independently of the Swift trigger. The target system was designed as two sites separated by ~ 100 km with 16 detectors on four mounts per site. Currently, there are two PIOTS installations, one at Mazagon, Spain, and one at the San Pedro de Atacama Observatory, Chile.

3.2.7 MASTER

The main goal of the MASTER project (Mobile Astronomical System of Telescope Robots) launched in 2002 in Russia is to produce a nightly survey of the full sky down to a limiting magnitude 19–20 (Lipunov et al. 2010). This photometric record will make it possible to address a broad range of astrophysical problems including supernovae and the nature of dark energy, exoplanet searches, gravitational microlensing, and solar system science. All MASTER-Net telescopes also respond to external alerts, primarily GCN. As of 2010, the network consists of three sites equipped with two types of instruments. MASTER VWF (Very Wide Field) based on 50- and 85-mm $f/1.4$ Nikor lenses covers up to $1,000 \text{ deg}^2$ down to ~ 12 mag. MASTER-II telescopes consist of a pair of 0.4-m catadioptric Maksutov tubes fitted with a $4K \times 4K$ pixel Alta U16 CCD cameras from Apogee and carried on a common German equatorial mount (► Fig. 2-14). The slew rate is up to 30 deg/s and the survey rate approaches $480 \text{ deg}^2/\text{h}$ assuming 1-min exposures (19 mag limit). MASTER-Net extends over 7 h in longitude and provides virtually continuous 24-h sky coverage during winter months. Over the past few years, MASTER telescopes detected a number of optical flashes from GRBs, discovered several supernovae, and performed synchronous observations of gamma-ray burst error boxes. Expansion plans include a total of seven nodes and new hardware to obtain multicolor photometric and polarimetric follow-up.

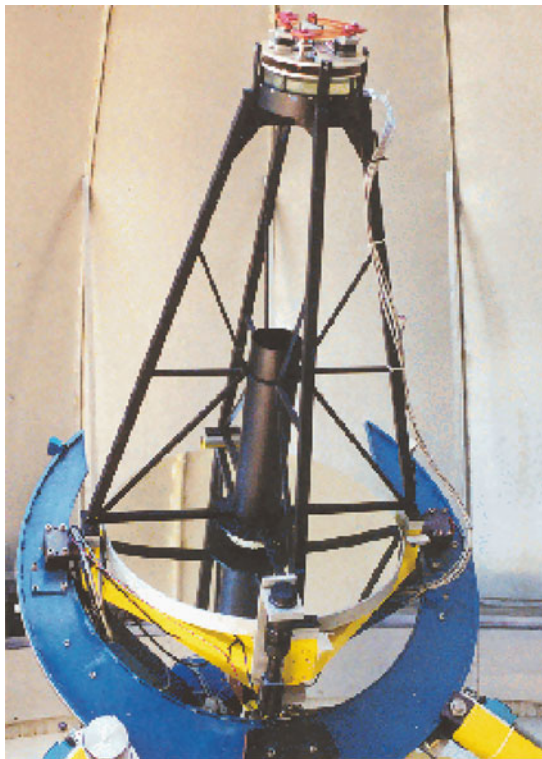


■ Fig. 2-14

MASTER-II robotic telescopes at Sternberg Astronomical Institute Caucas Observatory, Kislovodsk (Russia)

3.2.8 KAIT

The Katzman Automatic Imaging Telescope (KAIT) located at the Lick Observatory is dedicated to the search for optical transients and monitoring of celestial objects with a strong focus on supernovae (Filippenko et al. 2001). Its history goes back to 1990s and the Berkeley Automatic Imaging Telescope (BAIT; Richmond et al. 1993). The telescope is equipped with a $f/8.2$ Ritchey-Chretien mirror set on a light weight frame (● Fig. 2-15). Over nearly 15 years of operation, KAIT has been fitted with several different thermoelectrically cooled CCD cameras from Apogee and Finger Lakes. The pixel scale is 0.8 arcsec and the field of view is relatively narrow 6.7×6.7 arc min. This is well matched for GRB follow-up and pointed observations of galaxies to check for the presence of a supernova. Every 3–7 days, KAIT visits the same galaxies from a preprogrammed list and takes 16–20-s unfiltered exposure that can reach 19th magnitude objects. New images are automatically compared with reference templates, and candidate supernovae are then examined by observers. The best candidates are reobserved and upon confirmation released via International Astronomical Union Circulars (IAUC) and Central Bureau Electronic Telegrams (CBET). The emphasis is on catching and monitoring nearby supernovae



■ Fig. 2-15

The Katzman Automatic Imaging Telescope (KAIT), a 30-in. robotic telescope at the UCO Lick. Photo credit: Weidong Li and Alex Filippenko, University of California, Berkeley

in the Hubble flow before maximum light. Confirmed supernovae are automatically observed in *BVRI* filters on a 2–3 day cadence. Data taking is completely automated. The system checks the weather, opens the dome, points to targets, finds and acquires guide stars, takes exposures, and finally stores and processes the data, all without human intervention.

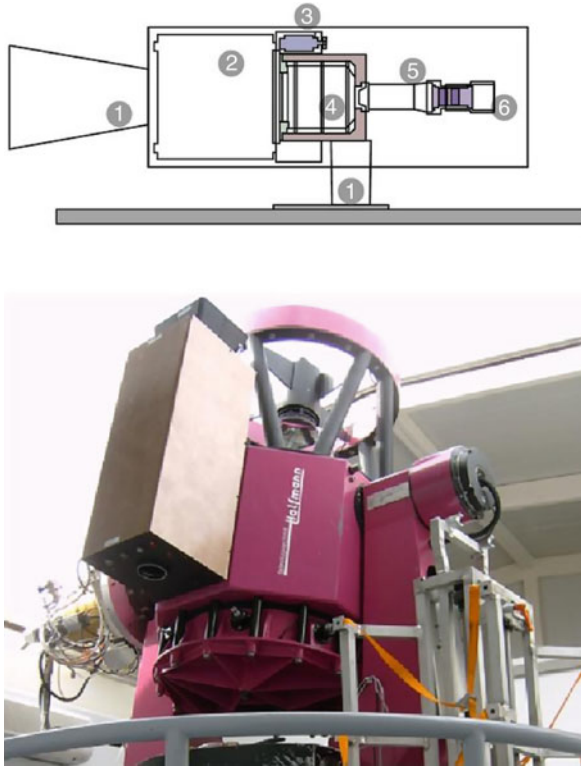
3.2.9 PAIRITEL

The Peters Automated Infrared Imaging Telescope (PAIRITEL) on Mt. Hopkins in Arizona is the first 1-m class IR instrument to achieve the status of a fully robotic system (Bloom et al. 2006). Until 2001, the 1.3-m telescope and simultaneous *J, H, K_s* imager were used by the 2MASS survey (● Sect. 3.3.2). After refurbishing work was completed in 2004, the new system began automated observing based on queue scheduling combined with an override mode for rapid response to targets of opportunity, primarily GRB localizations from GCN. Observatory software continuously monitors critical weather and control system diagnostics. Automated data processing pipeline handles the complexities of reducing IR images such as real-time dark current corrections or variable atmospheric transmission. Although PAIRITEL is not 100% autonomous, it became one of the most productive GRB follow-up instruments of the Swift era. Response times are typically ~2 min after receiving the alert and can be as low as 1.5 min, which is sufficient for catching some of the long duration bursts before the end of the gamma-ray emission. More recently, PAIRITEL also responds to triggers from optical surveys such as Palomar Transient Factory (PTF) and the Catalina Real-Time Transient Survey (CRTS).

3.2.10 REM, TORTORA, and MegaTORTORA

The main motivation behind REM (Rapid Eye Mount) was to enable prompt identifications of GRBs in the visual and near-IR wavelengths in order to study early afterglows, select high redshift events, and alert larger telescopes with spectrographs. REM is an Italian facility located at La Silla, ESO (Molinari et al. 2010). The telescope is a *f*/8 Ritchey-Chretien on a fast-slewing alt-azimuth mount that provides two stable Nasmyth focal stations suitable for fast motions (► Fig. 2-16). One of the Nasmyth foci hosts two instruments fed from a dichroic: ROSS visual camera and spectrograph and REMIR infrared imager with a filter wheel of ten slots including broadband *Z', J, H, K_s*, a narrowband H₂ filter, and a grism for low-resolution slitless spectroscopy. The mean time for response to a satellite trigger is 30 s. After a few years of semiautomated operation, the facility is now working autonomously and offers non-alert time to the astronomical community through regular proposal calls.

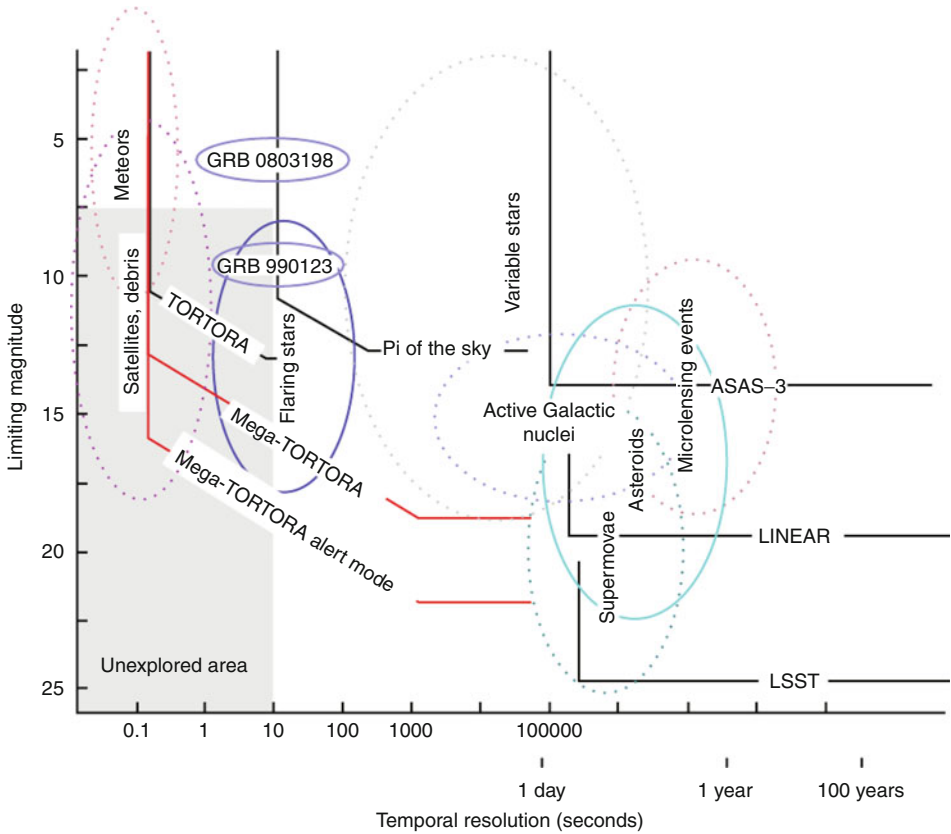
Since late 1990s, the Special Astrophysical Observatory (SAO) in Russia has been developing concepts and systems for optical sky monitoring with very high temporal resolution. TORTORA (Telescopio Ottimizzato per la Ricerca dei Transienti Ottici RAPidi) is a joined project between SAO and REM (Beskin et al. 2010). A schematic drawing of the instrument and the actual hardware attached to the REM telescope are shown in ► Fig. 2-16. The camera employs a TV-CCD detector and collects data at a rate of 7.5 frames per second with 0.128-s exposures separated by gaps lasting only 0.005 s. The resulting data rate is 20 Mb/s, and therefore



■ Fig. 2-16

TORTOREM telescope tandem at the La Silla Observatory (ESO, Chile): schematic view of the TORTORA detector (*top*) and the actual instrument attached to the fast-slewing NIR telescope REM (*bottom*). The TORTORA system records optical transients with 0.13-s temporal resolution and consists of the following components: (1) baffle, (2) objective lens, (3) focusing unit, (4) image intensifier, (5) transmission optics and CCD focuser, and (6) fast low-noise TV-CCD (Adopted from Beskin et al. (2010))

raw frames must be discarded after 1–2 days assuming a 1 TB storage buffer. Untriggered selection and classification of transients is performed automatically in real time by custom software. The TORTOREM strategy paid off with a spectacular detection of GRB 080319B (known as the “naked-eye burst”) and extremely high fidelity optical light curve simultaneous with gamma-ray emission. Plans for expansion involve increasing the sensitivity by 2–3 mag while preserving the current time sampling and a wide field of view, as well as new instrumentation for multicolor photometry and polarimetry. MegaTORTORA is envisioned as a collection of small telescope arrays with each telescope on a given mount observing the same field of view in a fixed photometric band and polarization state. Fast frame rates can be achieved using relatively affordable EMCCD cameras such as the one shown in ► Fig. 2-3. This new system is designed to push the envelope of the observable time scale and survey depth that can be reached with existing systems (◀ Fig. 2-17).



■ Fig. 2-17

Discovery space for fast celestial transients with optical emission. The ovals denote astrophysical phenomena, and the lines compare the capability of several instruments, both existing and planned for the future (Adopted from Beskin et al. (2010))

3.3 Dedicated Survey Telescopes

Specialized telescopes with moderate apertures in the range 0.5–3.0-m are often built for time-consuming survey work. Such instruments are typically optimized for a particular task and highly automated. The following sections provide examples of high impact projects in this category.

3.3.1 Microlensing Searches: OGLE and MOA

The first searches for gravitational microlensing in the Local Group of galaxies came online in the early 1990s. Three groups, EROS, MACHO, and OGLE, embarked upon then an enormous task of monitoring tens of millions of stars in extremely dense stellar fields of the galactic bulge and Magellanic Clouds searching for dark matter in the form of compact objects. This is necessary because the probability that a random star is microlensed at any given time is only one per

million. Today, OGLE (Optical Gravitational Lensing Experiment) and MOA (Microlensing Observations in Astrophysics) are continuing this quest.

Since 1996, OGLE is conducted using the 1.3-m Warsaw University Telescope at Las Campanas (Chile) equipped with successively larger detectors (Udalski et al. 1997). The telescope is a $f/9.2$ ($f/2.8$ primary) Ritchey-Chretien system of ultralow expansion (ULE) glass mirrors with a three-element field corrector giving a 1.5-deg diffraction limited field of view. The parabolic fork mount with friction drives provides precise control of the tracking rate in both RA and DEC, an essential feature for the OGLE-II phase of the project that collected data in the drift-scan mode without autoguiding. The telescope and instruments are operated from a control building located 15 m away from the dome. Survey operations are highly automated with only an occasional need for human assistance for data quality assurance. Real-time data reduction and analysis pipelines automatically process the incoming images and resulting light curves. Microlensing is in principle achromatic, and therefore most observations are conducted in a single band (I) with a small fraction of time invested in gathering basic color information (primarily V). The OGLE-IV experiment that began in 2009 employs the “third generation” camera based on a mosaic of 32 thin E2V $2,048 \times 4,096$ pixel CCD chips (over 250 million pixels) covering practically the entire usable area of the focal plane. A single “image” takes 20 s to read out and requires 0.5 GB of storage space. This latest upgrade constitutes almost an order of magnitude increase in survey capability compared to OGLE-III and is expected to increase the rate of microlensing detections to several thousand per year. Over 18 years of operation, OGLE produced an impressive stream of high-quality data and scientific results, most of them actually unrelated to microlensing. OGLE was the first microlensing survey to implement an alert system for reporting newly found microlensing events and light curve anomalies in near real time. The OGLE Early Warning System (EWS) dating back to 1994 is one of the earliest examples of this type of service (Udalski 2003). The project web pages offer public access to OGLE databases and other tools, including EWS.

Following a successful MOA-1 project, MOA-2 has been launched in 2004 on a dedicated 1.8-m telescope at Mount John University Observatory, New Zealand (Hearnshaw et al. 2006; Sumi 2010). The optics is characterized by a fast focal ratio at $f/3$. Corrector lenses have been designed to realize good image quality over the wide focal plane at prime focus. The current generation MOA-cam3 camera consists of ten E2V $2K \times 4K$ pixel CCD chips and covers a 2.2 deg^2 of sky in a single exposure. The camera is based on the GenIII system from Astronomical Research Cameras Inc. and can read the entire focal plane array in 25 s. A special wideband red filter was designed for a more sensitive microlensing search. General astronomical studies are mainly conducted using Bessell V , I filters. Thanks to a wide field of view, MOA-2 monitors 24 deg^2 every 2 h in the LMC and 50 deg^2 every hour in the galactic bulge. Faster cadence boosts detection statistics for brief planetary events on top of regular microlensing. In 2008, the combined yield of OGLE and MOA was over 1,000 microlensing events. Both projects are currently focusing on early detection of light curve anomalies that signal the presence of planets around stars acting as microlenses.

3.3.2 Near-Infrared Surveys: 2MASS, VISTA, and SASIR

The current “gold standard” of the NIR sky is the Two Micron All-Sky Survey (2MASS) (Skrutskie et al. 2006). Between 1997 and 2001, the survey collected over four million images covering 99.998% of the sky and obtained JHK_s photometry for 471 billion point sources with

S/N ratio of about 10 at 1 mJy and 1.6 million extended sources. To complete this monumental task, 2MASS used two highly automated 1.3-m telescopes: one at Mt. Hopkins, AZ, and one at CTIO, Chile. Each telescope was equipped with a camera capable of observing the sky simultaneously in all three bands separated by two dichroic mirrors and feeding a single NICMOS-3 256×256 pixel HgCdTe detector per channel. The secondary mirror had to be rotated smoothly during every 1.3 s exposure to “freeze” the field of view as the telescopes uniformly scanned the entire sky in declination at a rate of 57 arcsec/s. After each exposure, the secondary flew back to its starting position shifting the 8.5×8.5 arcmin field of view by about 1/6 of the frame size and producing six independent images of each object over the full scan. This brings the effective integration time to 7.8 s and allows sub-pixel “dithering” to enhance spatial resolution. The dead time for resetting the detector array and the secondary was less than 0.1 s. The overall efficiency of the survey was about 84%. In many ways, 2MASS is an exemplary survey project where both hardware and software sides of the house worked together to produce timely data releases that reinvigorated many fields of research.

VISTA, the Visible and Infrared Survey Telescope for Astronomy, is a 4.1-m, state-of-the-art wide-field near-IR survey facility at Cerro Paranal Observatory, ESO (Emerson and Sutherland 2010; Dalton et al. 2006). Its 3-m long, 67 megapixel cryogenic camera weighs 3 tons and covers a 1.65-deg field of view at resolution 0.34 arcsec/pixel. The focal plane is sparsely sampled with 16 VIRGO $2K \times 2K$ pixel HgCdTe detectors. Although the original specification called for *J*, *H*, and *K_s* filters, the detector QE remains at 70–80% down to 0.85μ . Thus, additional *Z* and *Y* filters were placed in the filter wheel to extend the science capability of the instrument. Active control of the position and tilt of the secondary mirror and curvature sensors inside the camera deliver excellent image quality over a wide field of view (0.66 arcsec FWHM). The cold-baffle design that minimizes the background combined with high optical throughput and detector QE allows VISTA to reach $J = 20.2$ mag objects in a 60-s exposure. Since the end of 2009, VISTA has been carrying out its original program of large-scale surveys of the southern sky that cut across a wide variety of science goals.

While there are many scientific questions that can be answered using deep, multi-epoch photometry in the infrared, the NIR time domain remains somewhat of a blind spot. The Synoptic All-Sky Infrared Survey (SASIR) is a proposal by the University of California at Berkeley and several institutions in Mexico intended to close this gap (Bloom et al. 2009). The proposed 6.5-m telescope at San Pedro Mártir Observatory in Mexico would perform simultaneous imaging in *Y*, *J*, *H*, *K* bands and repeatedly survey the entire sky north of $\delta = -30^\circ$ for 4 years to collect data on variability on time scales from minutes to months. The SASIR camera featuring $124 \times 2K \times 2K$ pixel arrays would be the largest IR imager ever constructed and would reach 100 times deeper than 2MASS (million times larger volume). “Etendue-couleur” (etendue times the number of photometric bands) for this system is more than three orders of magnitude larger than 2MASS and ten times that of VISTA. SASIR will for the first time reveal the dynamic infrared universe, study transients burried by dust, and discover thousands of quasars at $z \sim 7$ reaching to $z > 10$ to probe the epoch of reionization.

3.3.3 Moving Object Searches: Spacewatch, LINEAR, and CSS

Prior to 1980s, for almost a century, systematic searches for small solar system bodies were based on photographic methods. In 1998, NASA set out to meet the US Congressional mandate and discover 90% of all bodies larger than 1 km with potential for Earth impact by 2008. This greatly stimulated efforts to search for near-Earth objects (NEOs).

Spacewatch is the longest running search for moving celestial objects and was the first system to use CCDs to find new comets and asteroids (Gehrels 1991; Carusi et al. 1994). Along the way, the project made more groundbreaking contributions to survey astronomy, including the first use of drift-scan technique in astronomy, the first real-time software for moving object detection, and the first automatic discoveries of new moving objects of various types. The Spacewatch program at the University of Arizona's Lunar and Planetary Laboratory utilizes about 20 nights per lunation on Steward Observatory 0.9-m telescope and since 2002 also on the Spacewatch 1.8-m telescope on Kitt Peak, Arizona. Both telescopes scan the sky using modern CCD detectors and routinely discover new asteroids down to $V = 21\text{--}22$ mag.

The Lincoln Laboratory's Near-Earth Asteroid Research (LINEAR) program is a joined project between US Air Force, NASA, and MIT (Stokes et al. 2000). The project uses a pair of 1-m, $f/2.2$ robotic telescopes located at White Sands Missile Range near Socorro, New Mexico. Each telescope is fitted with a low-noise, frame-transfer $1,960 \times 2,560$ pixel thinned CCD designed and fabricated by MIT Lincoln Labs. The field of view is 2 deg^2 and allows covering the entire available sky each lunation down to ~ 19.2 mag with 10-s integrations. The LINEAR moving object detection pipeline requires at least three detections in a sequence of five visits separated by 30 min. At the end of 2007, LINEAR has discovered over 200 thousand new objects including 2,000 NEOs and 200 comets.

Since 2005, the Catalina Sky Survey (CSS) and its companion Siding Spring Survey (SSS) (Larson et al. 1999) dominate the rate of asteroid detections and new discoveries. The survey utilizes three telescopes: 1.5 m, $f/2$ telescope on Mt. Lemmon; 0.68-m, $f/1.7$ Schmidt telescope near Mt. Bigelow (both near Tucson, Arizona); and a 0.5-m $f/3$ Uppsala Schmidt telescope at Siding Spring Observatory in Australia. The cameras are cooled to approximately -100°C , so their dark current is about 1 electron per hour. Each telescope is equipped with a $4,096 \times 4,096$ pixel camera that covers a field of view between 1 and 9 deg^2 depending on the telescope. The 1.5-m instrument can reach a detection limit of $V = 21.5$ mag in 30 s. CSS discovers 300–600 NEOs per year and has a distinct advantage of reaching the southern sky. Over the past few years, the Catalina survey also became a major supplier of real-time detections of transient photometric variability.

3.3.4 Spectroscopic Surveys

Spectra are like astronomical fingerprints that immediately reveal the physical nature of the object at hand. But collecting a large number of them in a reasonable time is not easy. Spectroscopy requires large light collecting areas and/or long integration times by comparison with photometry, automated target acquisition is tricky, and special measurement techniques must be used to handle the added dimension along the wavelength. Fortunately, recent advances in photonics and robotics are making it possible to take hundreds or even thousands of spectra at the same time and automate most if not all of the work.


Several existing telescopes have been fitted with new generation multi-object spectrographs and harnessed by spectroscopic surveys. The two-degree field facility (2dF) on AAT uses flexible optical fibers to collect light from 400 objects located anywhere within a 2-deg field of view (Lewis et al. 2002). This light is then directed to a spectrograph that records all 400 spectra simultaneously with a CCD. Other examples of ongoing or completed spectroscopic surveys are RAVE using the 6dF instrument on UK Schmidt at Siding Springs (Steinmetz 2003) and a suite of SDSS-II/III surveys (SEGUE, APOGEE, MARVELS, BOSS) on Apache Point 2.5-m telescope (Eisenstein et al. 2011). A necessarily incomplete list of planned spectroscopic

survey projects includes HERMES on AAT, BigBOSS on Mayall, WINERED in Japan, and LAMOST in China. The BigBOSS instrument features 5,000 robotically actuated fibers covering a 3-deg-diameter field of view and feeding ten identical spectrographs each covering the wavelength range from 340 to 1,060 nm with a resolution $R = 3,000\text{--}4,800$. LAMOST is a 4-m Schmidt telescope in a meridian-crossing configuration with 4,000 fibers over a 5 deg field. The Hobby-Eberly Telescope (HET) (Hill et al. 2008), a 9.2-m telescope located at the McDonald Observatory, and its twin SALT in South Africa are primarily spectroscopic facilities built according to a new revolutionary design that dramatically cuts the construction cost. The telescope stays at a fixed elevation above the horizon and rotates in azimuth to access most of the useful sky.

Completely autonomous medium to high-resolution spectroscopy on 1- and 2-m class telescopes is now feasible (Plokhotnichenko et al. 2010). STELLA is a fully robotic observatory hosting two 1.2-m altazimuth telescopes located on Tenerife, Spain (Strassmeier et al. 2010). Stella Echelle Spectrograph (SES) delivers spectra between 390 and 860 nm with resolving power up to 75,000 recorded on a single CCD frame. The autoguiding unit must position the target on a diaphragm with sub-arcsecond precision and maintain the alignment for an hour. A beam splitter sends 4% of the target light to a guiding CCD. Off-axis light is redirected to form another image of the target. When the alignment is perfect, both images coincide. The design allows adjusting the telescope focus and guiding during the same readout cycle. While STELLA is primarily devoted to studies of stellar activity, it is also used for other programs, including rapid follow-up of targets of opportunity. Several groups are developing systems for integrated field unit (IFU) spectroscopy for rapid response telescopes around the world that can continuously follow transients brighter than ~ 22 mag. An example is the Las Cumbres Observatory Global Telescope Network of 2-m telescopes carrying photometric/IFU instruments dedicated to follow-up (Hidas et al. 2008).

3.4 Large Robotic Telescopes

3.4.1 Liverpool Telescope

The Liverpool Telescope is a 2.0-m fully robotic telescope located at the Roque de Los Muchachos Observatory on La Palma, Canary Islands (Steele et al. 2004). It is owned and operated by the Astrophysics Research Institute of Liverpool John Moores University (JMU) in England. While most robotic telescopes are strongly dominated by a single science program, the objective of the Liverpool Telescope is to provide a generic research facility offering a broad suite of instrumentation and allocated by independent time allocation panels. Routine nighttime operations of the Liverpool Telescope are completely autonomous, including configuration of instruments, selection of targets, and scheduling observations while adapting to changing weather conditions (Smith et al. 2010). This is a truly impressive feat for a 2-m class telescope (shown in  Fig. 2-18). Since 2010, programs approved by the allocation committee can be uploaded by observers directly into the scheduling database at any time during the day or at night for autonomous execution by the system. The observatory is particularly well suited for ambitious programs requiring robotic response to unpredictable events and their follow-up, monitoring variable sources on time scales from seconds to years, and coordinated simultaneous observations with other facilities in space and on the ground. Together with the Faulkes



■ Fig. 2-18

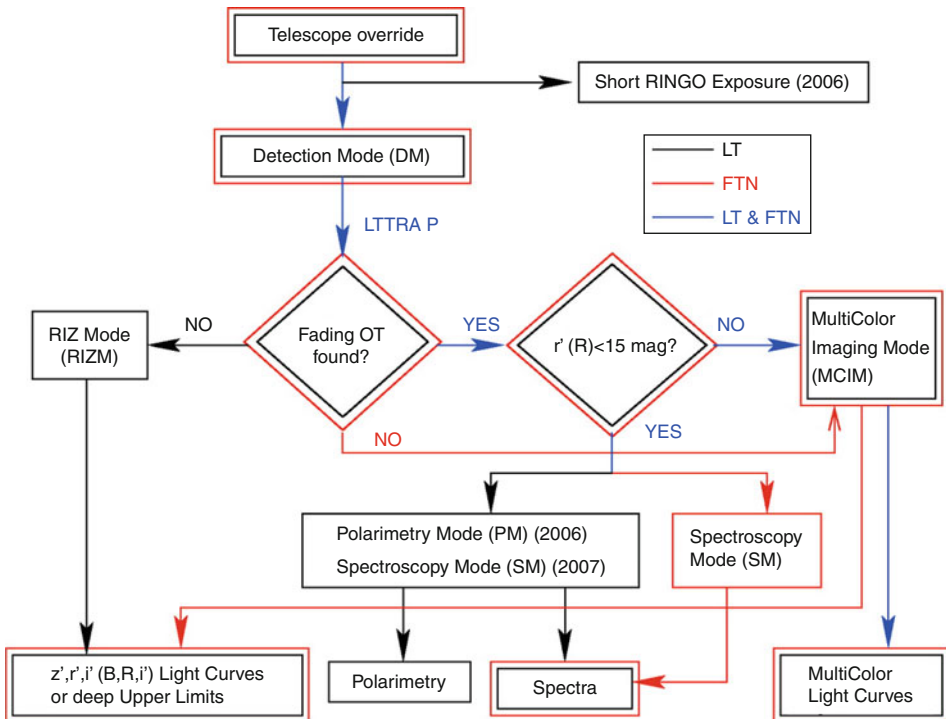
Liverpool Telescope and its founder professor Mike Bode. This 2-m giant in the world of robotic telescopes is capable of observing transient events at random locations within minutes of the localization alert using one of the five permanently mounted instruments

Telescope North (FTN) in Maui, Hawaii, and the Faulkes Telescope South (FTS) in Siding Spring, Australia, the Liverpool Telescope (LT) forms a global network of 2-m telescopes called RoboNet (Tsapras et al. 2009).

The LT was designed and constructed by Telescope Technologies Limited (TTL), a spin-off company of JMU. The design is Ritchey-Chretien/Cassegrain at $f/10$ placed on an altazimuth mount. Up to five instruments can be permanently mounted on the telescope. An instrument is selected for observing using a rotating tertiary mirror in the optical path. As of 2010, the LT carried the following instruments: (1) optical imager RATCam, (2) ultrafast optical polarimeter RINGO2, (3) infrared imager SupIRCam, (4) intermediate dispersion dual-beam spectrometer FRODO-Spec, and (5) fast-readout optical camera RISE. Perhaps the most interesting on this list in terms of innovation is the RINGO polarimeter (Steele et al. 2010) capable of measuring polarization of highly variable sources. The idea is to use a fast rotating Polaroid to modulate the polarized part of the flux and deflect the light using a wedge prism, also rotating. This way, a polarization signal forms a $\sin 2\theta$ pattern around a ring. A serious disadvantage of this original design is spreading of the signal over many CCD pixels compounded by source confusion due to overlapping rings of neighboring objects. RINGO2 addressed these shortcomings

by eliminating the prism and recording the fast variable signal directly using an electron multiplication CCD. The current system employs the Andor iXon+897 camera hosting a 512×512 E2V CCD97 detector in a Peltier cooled package. The Polaroid rotates at 60 RPM, and the signal is divided into eight time bins per rotation resulting in 125 ms exposures, well within the specification of the EMCCD device.

The unique capability of LT can be illustrated using real-time GRB follow-up. [Figure 2-19](#) shows the flowchart of LT-TRAP, the Liverpool Telescope Transient Rapid Analysis Pipeline (Guidorzi et al. 2006). Upon receiving a GRB localization alert from a gamma-ray observatory via a live GCN socket, the telescope enters the override mode, automatically slews to the target position, and within 2–3 min of the GRB alert begins observing. During the first phase of the response (detection mode), two sets of three 10 s exposures using the r' filter are taken. The data is automatically reduced and analysed. A variable object in the error box is interpreted as the candidate optical counterpart. At this point, LT-TRAP determines the most productive observing strategy given what has been seen (or not seen) so far and puts the telescope in one of the detailed follow-up modes: multicolor imaging, polarimetry, spectroscopy, or a series of deeper exposures using the Sloan $r' i' z'$ filters. Among other interesting results, the LT-TRAP system delivered the first measurements of polarization in the early optical GRB afterglow and showed that the level of polarization can be as high as 10%. In addition to rapid GRB follow-up,



■ Fig. 2-19

Flowchart of the LTTRAP observing and data analysis strategy used at the Liverpool Telescope and the Faulkes Telescope North facilities. This sophisticated software system is designed to optimize the rapid response sequence following GRB localization alerts (Adopted from Guidorzi et al. (2006))

the list of science programs on LT includes observations of extrasolar planets detected using both the transit method and gravitational microlensing anomalies, monitoring of active galactic nuclei, and follow-up of other explosive transients such as cataclysmic variables and supernovae.

3.4.2 GROND

GROND (Gamma-Ray Burst Optical and Near-Infrared Detector) is a seven-channel imager specifically designed for GRB afterglow observations (Greiner et al. 2008). High-resolution spectroscopy is the key probe of detailed physical conditions in the burst environment and must be obtained within the first few hours after the explosion. High redshift events are particularly important and best identified with multiband photometry in optical/NIR providing a photometric redshift based on the Ly α break. The initial positions of GRBs are only accurate to a few arcminutes, so the associated transient must be first identified using time-resolved images covering the entire error box. Simultaneous multi-color images provide robust measurements of the spectral energy distribution (SED) for an afterglow that fades by 2–3 mag within the first few minutes. A 2-m class telescope is required to detect a large fraction of GRB afterglows. GROND is a state-of-the-art instrument designed to address all those issues.

GROND is installed on the ESO's 2.2-m MPI telescope on La Silla (Chile). The telescope is a f/8 Ritchey-Chretien with a focal length of 17.6 m on an equatorial fork mount. The instrument employs a focal reducer in the infrared to produce a 10×10 arcmin field of view and delivers $0.4''$ images (FWHM). Different photometric bands are separated using multiple dichroics that reflect the short-wavelength light, while the long wavelengths pass through. GROND resides on the side of the telescope in a Coude-like focus. During a GRB alert, a movable flat mirror is folded in the main telescope beam to direct the light into GROND. The visual channels utilize $2,048 \times 2,048$ pixel back-illuminated E2V devices with $13.5\text{-}\mu$ pixels. The field of view in *griz* is 5.4×5.4 arcmin and determined by the telescope focal length and the size of the CCD. In the *JHK*, a $1,024 \times 1,024$ pixel Rockwell HAWAII-1 array combined with the focal reducer delivers images covering 10×10 arcmin. The entire assembly resides in a single cylindrical cryostat. A two-stage closed-cycle coolers (CCC) maintains different operating temperatures for the visual and NIR benches (80 and 65 K, correspondingly).

It would be difficult to take advantage of GROND's capabilities without a sophisticated control software that coordinates exposure and readout cycles of various components of the system. The control system is integrated with the ESO scheduling using standard observation blocks to facilitate rapid response and coordinate the interaction of GROND with other instruments on the 2.2-m telescope. The GROND Pipeline is a software package written specifically for automated scheduling and analysis of GROND observations. The system consists of multiple processes that work asynchronously. The processes in the observation control layer receive GRB alerts from the GCN connection, calculate the visibility, as well as schedule and interrupt/reschedule observations. Unlike the Liverpool Telescope, GROND does not use JIT (just in time) scheduling but rather plans ahead to observe all GRBs that are accessible and makes the schedule known to observers working with other instruments several hours in advance. Conflicts are resolved based on a priority scheme. Once the observation is accepted, the system executes a standard rapid response mode procedure adopted from the VLT and creates a master process that controls the analysis of data from all seven channels collected from all relevant observation blocks. Subsequent data analysis depends on the user-selected strategy. There are 22 different arrangements of data analysis strategies mapped onto 12 processes. The full data processing chain consists of (1) automated data reduction to obtain astrometry and

time-resolved multicolor photometry for all objects in the field, (2) comparisons with external catalogs and GRB data to establish the best candidates for the optical GRB counterpart, and (3) testing for the presence of the Ly α break and derivation of the photometric redshift using the public HyperZ code. Since its commissioning in April 2007, GROND has enabled many important contributions to GRB science, including rapid redshift determination for GRB 090423 at $z = 8$, the most distant object ever observed.

3.5 Large-Scale Surveys

3.5.1 SDSS

The Sloan Digital Sky Survey (SDSS) is arguably the most ambitious and influential astronomical survey completed to date (York et al. 2000). The survey was funded and carried out by a consortium of 25 institutions from around the globe. Using a dedicated telescope at Apache Point Observatory in New Mexico equipped with state-of-the-art instruments, the core projects SDSS-I and SDSS-II collected approximately 5 terabytes of deep multicolor images covering more than a quarter of the sky, high precision photometry for 230 million celestial objects, and more than one million spectra. Special purpose SDSS-III programs will continue through 2014. This extraordinary observational material is available online to anyone willing to use it and has literally transformed the astrophysical research.

The SDSS telescope (Gunn et al. 2006) is a 2.5-m $f/5$ modified Ritchey-Chretien altitude-azimuth telescope. A large secondary mirror and two corrector lenses provide a large distortion-free field of view. In order to maximize the efficiency of the survey, SDSS imaging was collected by drift-scanning along great circles at the sidereal rate and adjusting the CCD readout accordingly. The data is then divided into manageable pieces for processing. The telescope was equipped with the multi-color photometric camera and two multi-fiber spectrographs. Besides the 2.5-m telescope, the SDSS made use of the photometric calibration telescope, a seeing monitor, and a cloud scanner.

The SDSS imaging camera (Gunn et al. 1998) contained two sets of CCD arrays: the imaging array and the astrometric arrays. The imaging array consists of 30 $2,048 \times 2,048$ pixel Tektronix CCDs with filters placed directly in front of them. The pixel size is 24μ (0.36 arcsec on the sky). Detectors are arranged in six columns and five rows. Each row observed the sky in one color, and the direction of the scan was along the columns. As a result, each point on the sky passed through all filters in a temporal sequence $u'g'r'i'z'$. Each CCD covered a swath of sky 13.5 arcmin in width. The gaps between the columns required two scans to produce a filled stripe 2.54 deg wide. The camera design allowed for about 8% overlap between the two “half-scans” on each side to ensure that no area was lost. The resulting integration time was 54.1 s per filter. The northern galactic cap was covered by 45 such great-circle arcs. In many ways, SDSS served as a path-finder mission for large-scale surveys in the era of electronic imaging and established the standard of quality for future time domain surveys.

3.5.2 Palomar Transient Factory

The Palomar Transient Factory (PTF) is an automated, wide-field survey dedicated to a systematic exploration of the transient optical sky (Law et al. 2009). To the extent possible, existing

technology was reused to lower the cost with modifications necessary for achieving the science goals. The survey is performed using the Samuel Oschin telescope at Palomar Observatory, a 48-in. Schmidt with a glass corrector plate, and a 72-in. primary mirror at $f/2.5$. The same telescope was used for Palomar Sky Surveys (POSS-I/II) and Palomar-QUEST digital synoptic sky survey (Djorgovski et al. 2008).

Changes to the optics included replacement of the flat dewar window with a field corrector designed to handle the curvature of the focal plane, tuning primary mirror supports, refurbishing of optical surfaces, improved baffling, and installation of internal calibration lights. The original LN cooling system was replaced with a closed-cycle system based on a Polycold Compact Cooler with no moving parts to minimize vibration. A commercial shutter from Scientific Instrument Technology provides fine control of the exposure time to within 2 ms with minimal vignetting. Filter exchanges are completed in 15 s using a microstepper motor and a custom-made mechanism that selects one of two filters preselected for a given night.

The CCD detector is based on the CFH12K design developed for the prime focus of the Canada-France-Hawaii Telescope (CFHT). It is a $12\text{K} \times 8\text{K}$ pixel mosaic of $6 \times 2\text{K} \times 4\text{K}$ pixel MIT/LL CCID20 CCDs. The array readout time was lowered to 30.7 s (about 50% relative to the original design) to match the telescope slew and settle time for PTF operations. The survey is conducted robotically by the Observatory Control System (OCS) written in MATLAB. OCS issues command to the camera, filter exchanger, shutter, telescope, dome, and focuser based on the feedback from the scheduler, the telescope/camera, a weather station, and the data quality monitor. Most PTF images (92%) are taken through either the SDSS g' filter or Mould- R in order to optimize the detection sensitivity in full moon and dark-sky conditions, respectively.

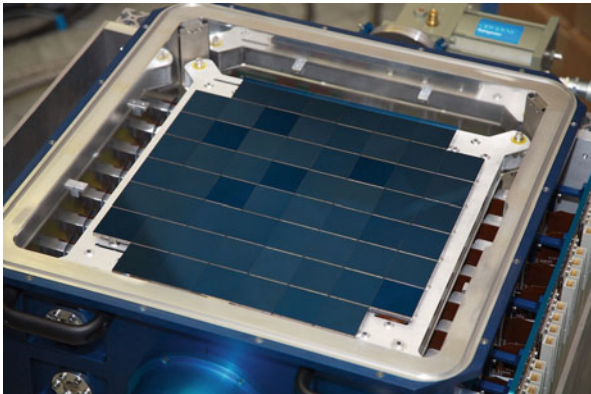
For maximum efficiency, four distinct experiments take turns on P48 throughout the year: (1) explosive transient search on a 5-day cadence, (2) dynamic cadence experiment designed to explore transient phenomena on time scales between 90 s and 1 day, (3) a fast-cadence search in the “stare mode” to discover eclipsing binaries and transiting planets in Orion, (4) a deep survey in $H\alpha$ covering 3π sr of sky for three nights each full Moon. PTF performs automatic real-time transient detection using difference images between program frames and historical template images. The Transients Classification Pipeline utilizes a Bayesian framework to provide machine generated type determination for detected transients. PTF conducts semiautomated follow-up of the most interesting transients, primarily using the automated Palomar 60 telescope, and also Palomar Hale Telescope, the Keck telescopes, PARITEL, and the Las Cumbres Observatory telescope network. Since the beginning of regular survey operations in the fall of 2009, PTF has been delivering a steady stream of transient detections, mostly newly discovered supernovae. PTF is an important milestone in the development of autonomous systems for detection, classification, and follow-up of astrophysical transients.

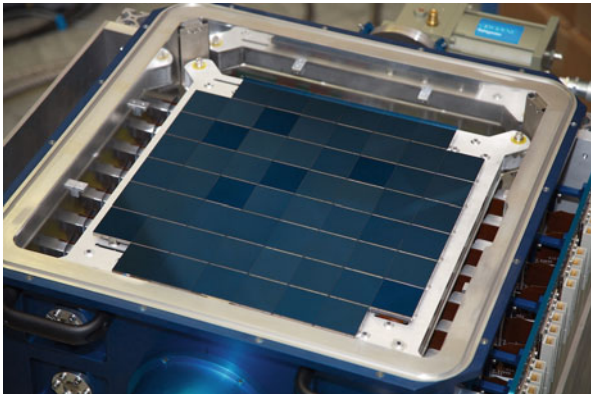
3.5.3 Pan-STARRS


The Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) project is an implementation of the synoptic survey concept (Kaiser et al. 2010). It has been funded by the US Air Force for the design and construction of the observatory facility with supporting control software, image processing pipeline, and database system. Survey operations are supported by a consortium of academic partners in the USA, Germany, the UK, and Taiwan and partially NASA. While addressing the threat associated with near-Earth objects (NEOs) is one of the original science drivers for the project, Pan-STARRS is optimized to tackle a wide

spectrum of problems related to the solar system science, time-domain astrophysics, galactic and extragalactic astronomy, and cosmology, including dark energy studies.

The target Pan-STARRS system, PS4, will be composed of four individual telescopes (Hodapp et al. 2004), each with a 1.8 m diameter mirror observing the same patch of sky simultaneously. This “distributed aperture” approach is more cost-effective and carries a much lower technological risk compared to a large monolithic mirror design with the same collecting area. The price for added flexibility such as the freedom to use four different filters at once, or temporarily point each telescope in a different direction, is four times more pixels and computing power per survey area. The importance of the latter factor is diminished by continuing exponential fall in the cost of electronics. Upon completion, the PS4 system will have an etendue of 60, i.e., five times better than any existing instrument. Optical layout is based on Ritchey-Chretien design with a Cassegrain focus and a three-lens corrector. The focal ratio of $f/4$ gives a focal length of 8 m and a plate scale of $38.5 \mu/\text{arcsec}$. Each telescope will have a 3-deg field of view and be equipped with a 1.4 billion pixel CCD camera. The spatial sampling of the sky will be about 0.3 arcsec. When searching for potential killer asteroids and variable objects, Pan-STARRS covers $6,000 \text{ deg}^2$ per night. The entire sky as seen from Hawaii will be observed three times during the dark time per lunation. Searches for asteroids and NEOs may utilize a very wide filter covering most of the visible waveband from 0.5 to 0.8μ . For other works, Pan-STARRS employs standard SDSS filters $g'r'i'z'$. Excellent near-infrared response of the Pan-STARRS detectors also allows the use of a y -band filter at 1μ .

The Pan-STARRS project has completed its first 1.4 gigapixel focal plane array, the world’s largest CCD camera (Onaka et al. 2008). The detector consists of 60 densely packed $4K \times 4K$ pixel MITLL CCD orthogonal transfer arrays (OTAs) developed at MIT Lincoln Labs (Tonry et al. 2008). The focal plane array is shown in  Fig. 2-20. Each OTA is an 8×8 array of 600×600 pixel CCDs known as “cells,” each equipped with its own logic and output. OTAs allow rapid shifting of the collected charge during exposure. This functionality is the basis of the sophisticated on-detector image compensation employed by Pan-STARRS, an electronic equivalent of the tip-tilt correction using tiny motions of the secondary mirror. A camera composed of OTAs automatically functions as a telescope guider. The readout is performed using




 Fig. 2-20

Focal plane of GPC1, the first gigapixel CCD camera of Pan-STARRS

480 outputs operating at a rate of 1 Mpixel/sec with gigabit Ethernet interfaces. The new controller electronics developed for this purpose called STARGRASP can be scaled to even larger focal planes.

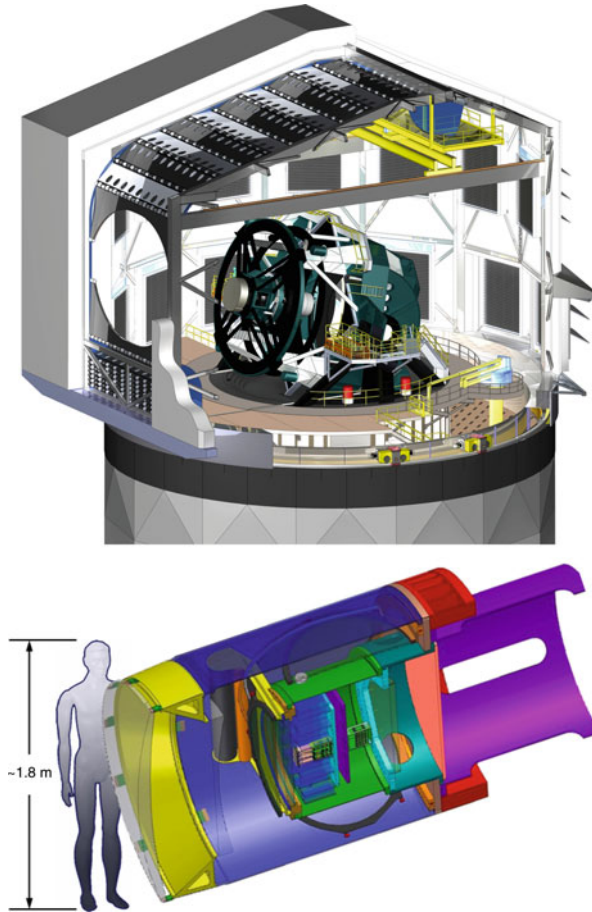
The first Pan-STARRS system, PS1, is located on Mount Haleakala and started regular observing in March 2009. During a “demonstration month” it produced an average of ~500 images per night and detected several hundred supernovae and thousands of asteroids. The PS1 science mission began on May 13, 2010. A single observation with a broadband filter will reach a 5σ depth of 24 mag. By adding observations taken over several years, Pan-STARRS should be able to reach a maximum depth of 29.4 mag.

3.5.4 LSST

The Large Synoptic Survey Telescope (LSST; LSST Science Collaborations et al. 2009) has been selected as the top priority large ground-based project by the Astro2010 Decadal Survey of the National Academies. The design grew out of ideas surrounding the Dark Matter Telescope, a large wide-field telescope with enormous etendue proposed around the turn of the century.  Figure 2-21 illustrates the concept. The effective etendue of LSST will be $319 \text{ m}^2 \text{ deg}^2$, more than an order of magnitude improvement over existing facilities. There are four major science drivers that shape the current reference specification for LSST (Ivezic et al. 2008): (1) taking an inventory of the Solar System, (2) mapping the Milky Way, (3) exploring the transient optical sky, and (4) probing dark energy and dark matter. Remarkably, the same instrument will enable breakthrough discoveries in all these areas without introducing imbalanced engineering requirements and multiple survey strategies. One can think of this as massively parallel astrophysics. Each patch of sky will be visited 1,000 times over 10 years, producing trillion measurements with temporal astrometric and photometric data on 20 billion objects. About 90% of the observing time will be devoted to a uniform deep-wide-fast survey mode. The survey will cover $30,000 \text{ deg}^2$ with south of declination $\delta = +34.5^\circ$, $20,000 \text{ deg}^2$ of which will be covered with the deep-wide-fast pointings. The remaining 10% of the observing time will be allocated to special programs such as fast-cadence monitoring of selected fields.

The LSST optical layout (Seppala 2002) is a modified Paul-Baker three-mirror system with a three-lens refractive field corrector. The color filter is placed in front of the third lens that also serves as the entrance window to the cryostat with the sensor at the focal plane. This design produces a large field of view with uniformly excellent image quality. The primary-tertiary mirror is 8.4 m in diameter, equivalent to a 6.7-m-diameter unobscured clear aperture. Overall, the optical system is extremely fast for its size with the f /ratio 1.23 at the effective focal length 10.3 m. The LSST camera assembly will be inserted through a 1.8-m opening in the secondary. The effective aperture diameter of 6.5 m was chosen to achieve the coadded survey depth $r \sim 27.5$ mag (5σ , point source), given the total exposure time of 30 s per visit (2×15 -s split) and the 10-year duration of the project. In a single visit, LSST will cover 9.6 deg^2 (3.5 deg FOV) down to $r \sim 24.5$.

The key to achieving a high duty cycle and minimizing the loss of exposure time is a short slew and settle time that is beyond the capability of today's large telescopes. The LSST telescope will be one of the most accurate and agile large telescopes ever built, despite its substantial size and weight. This is accomplished with an innovative design that exploits the small focal ratio of the telescope. The result is a very compact and stiff structure of welded and bolted steel on a powerful altitude over azimuth mount. A nominal move by 3.5 deg in elevation and 7 deg in azimuth will take only 5 s, including the time to acquire the new pointing with an error below



■ Fig. 2-21

Proposed Large Synoptic Survey Telescope (LSST): 8.4-m telescope in the enclosure (*top*) and 3.2-gigapixel, six-color camera (*bottom*)



0.1 arcsec. Allowing 2 min per filter change, LSST will spend 80% of the available time actually exposing the sky (weather permitting). The pointing accuracy across the sky is expected to be better than 0.2 arcsec.

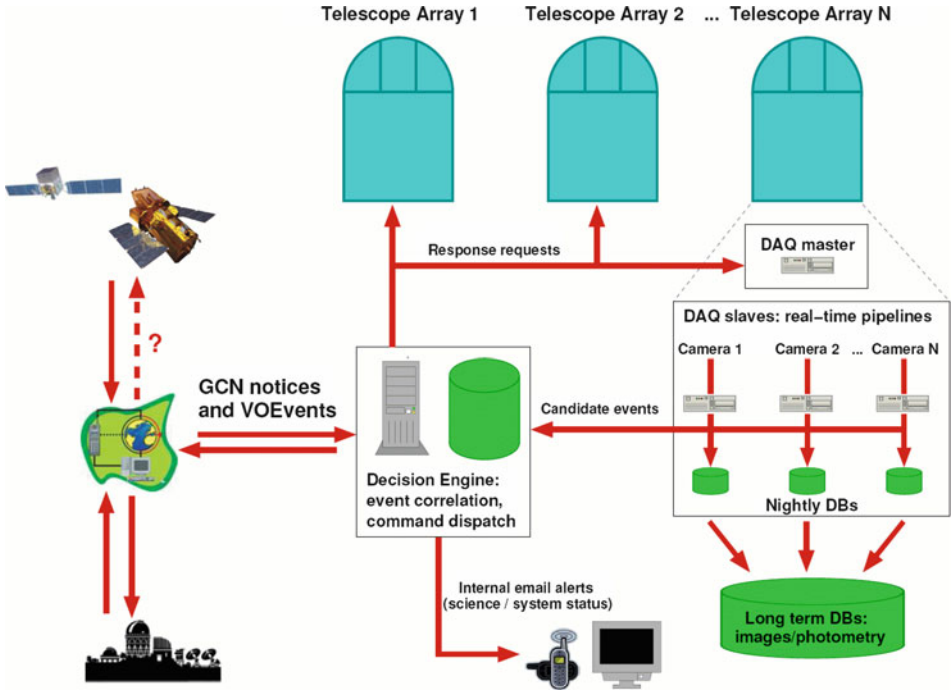
The LSST camera (Gilmore et al. 2008) features a 3.2-gigapixel focal plane array of 189 $4K \times 4K$ pixel, CCD sensors with $10.0\text{-}\mu$ pixels. At the plate scale of 0.2 arcsec/pixel this corresponds to an area 64 cm in diameter paved with silicon! The camera employs deep-depletion, back-illuminated CCD devices, each segmented into 16 channels to enable reading out the entire array in just 2 s. Individual detectors, as well as the associated camera electronics, are organized into relatively independent units called “rafts” hosting 3×3 CCDs and placed in a vacuum cryostat that maintains the operating temperature -100°C . The key characteristics of the camera that will make LSST science possible are (1) high quantum efficiency between 320 and 1,080 nm, (2) minimal PSF degradation due to charge diffusion, (3) very flat focal plane, (4) tightly packed design, and (5) fast readout enabled by highly segmented detectors.

LSST will employ active optics image corrections using wavefront measurements collected with special purpose CCDs mounted at the corners of the science array. The feedback loop that maintains the tracking of the telescope relies on the guide sensors that sample the locations of bright stars ten times every second. The camera body also contains a mechanical shutter and a sophisticated filter exchange mechanism holding five of the six LSST optical filters (*ugrizy*) arranged in a 3-D structure to minimize the obstruction in the optical path. The sixth filter can only be swapped in during daylight. The LSST filter system bears a strong similarity to the one used for SDSS (Fukugita et al. 1996). It covers the observed wavelength range with roughly logarithmic spacing and samples the Balmer break, while avoiding prominent telluric features. A notable extension is the presence of the *y* band due to high sensitivity of the deep-depletion CCDs at 1 μ . LSST filters are characterized by nearly perfect transmission, very flat peaks, and only a small amount of cross-talk between the bands. They were designed to achieve 0.5% relative photometric calibration.

4 Telescope Networking

Networking is a powerful way to combine resources. Computer networks have completely revolutionized the way we access and process information. They are also one of the major enabling technologies behind the proliferation of autonomous telescopes and observatories. Telescope networks are the next logical step in this progression and are already delivering previously impossible observations. Nevertheless, the full potential of linking telescopes, databases, and search engines for real-time knowledge extraction and observing optimization is yet to be unleashed. The field is undergoing rapid growth driven by the fusion of major technologies: robotic telescope hardware, cutting-edge storage devices, online databases, and state-of-the-art knowledge extraction algorithms that work together as an end-to-end system (Allan et al. 2006c).

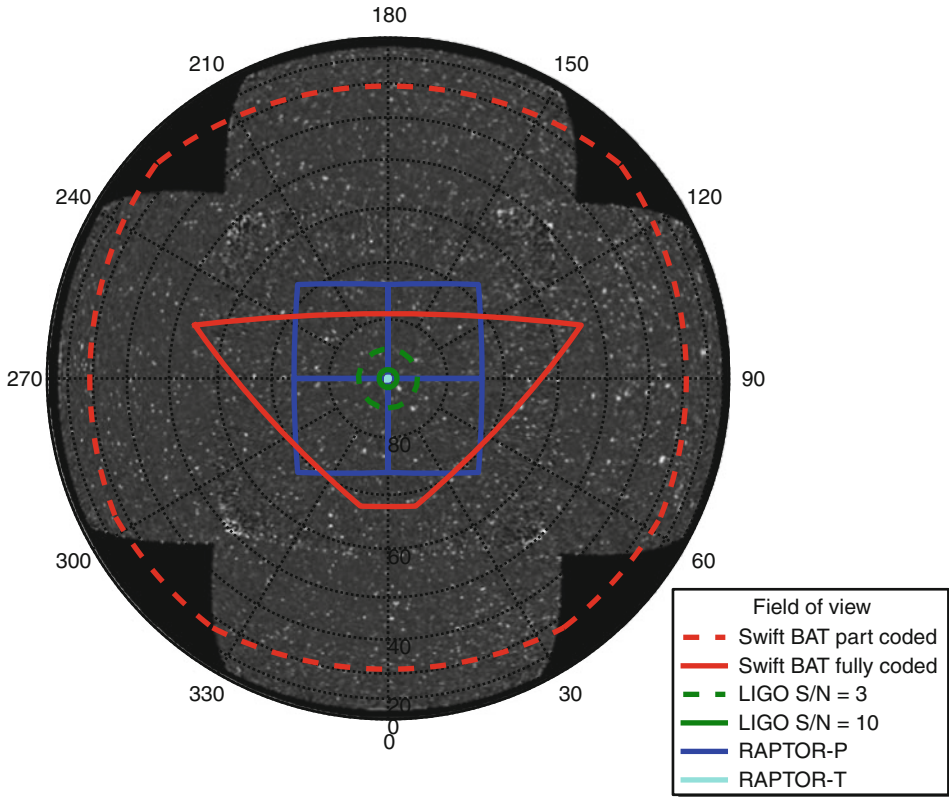
The concept of a telescope network is illustrated in  Fig. 2-22. A geographically distributed ensemble of survey instruments is systematically searching the sky for interesting events. Depending on the target science, each site may actually host an array of telescopes. The RAPTOR/Thinking Telescopes network for transient detection and monitoring is a good example of such system on a single project scale ( Sect. 3.2.5). The data is processed on the fly by a set of dedicated computers running image processing pipelines, performing astrometric and photometric calibrations, real-time transient recognition, and reporting of interesting changes to a central decision engine that determines subsequent actions. The system is integrated with databases that provide historical context information, both internal and external to the project. This information is used by the decision engine to “interpret” events and attempt inferences about their nature. Response requests are then sent automatically to follow-up instruments for the most promising candidate transients. Follow-up telescopes are typically more sensitive and provide detailed information such as multicolor photometry or spectra but require well-localized targets. The results of the follow-up are fed back to the decision engine. This autonomous “closed-loop” operation with little or no human supervision enables a radically new approach to exploring the variability of the night sky on short time scales. It will also play a crucial role in the Peta-scale surveys of the future such as LSST. One immediate consequence of this topology is that the distinction between survey and follow-up is slowly disappearing. The same instrument assumes different roles depending on what is happening at the moment.



■ Fig. 2-22

Schematic view of a modern telescope network. Note the presence of multiple feedback loops that determine new data acquisition activity based on what has been inferred so far

Local networks are integrated with the global network via event messaging protocols and the so-called aggregator nodes. Aggregators process information from a heterogeneous collection of instruments including ground-based telescopes and space observatories covering all parts of the electromagnetic spectrum or even non-electromagnetic signals (gravity waves and neutrino). With the new generation of sensitive neutrino observatories and gravitational wave detectors (Icecube, e/a-LIGO, and LISA in the future) coming online, we are witnessing the birth of multi-messenger astronomy. However, the astronomical community is still rather unprepared to take advantage of these developments. The event of the century could easily be missed if it occurred in the galactic plane or appeared as a saturated object in one of the deep wide-field surveys. Optical and NIR robotic observatories are helping to gradually fill these blind spots, but the level of coordination between the various survey and follow-up instruments must be increased to make this strategy effective (Stubbs 2008). ▶ [Figure 2-23](#) illustrates the difficulty involved in this type of work that requires cross-correlating and combining observations made with instruments of vastly different coverage and spatial resolution. The significance and the assigned level of priority of a random low signal-to-noise detection strongly depends on the available context information. The data stream from LIGO is full of possible 3σ detections that will normally be ignored, unless we can reliably link any of them to rapid optical variability. Telescope networking is about enabling this type of analysis on a global scale.



■ Fig. 2-23

Illustration of challenges involved in correlative multiwavelength, multi-messenger observations of cosmic explosions. An instantaneous view of the sky as seen by the RAPTOR-Q all-sky monitor (▶ Sect. 3.1.2) serves as the background. The field of view of the Swift/BAT gamma-ray telescope is compared to a typical error box associated with a gravitational wave detection from LIGO and the field of view covered by modern rapidly slewing optical telescopes, RAPTOR-P, and RAPTOR-T (▶ Sect. 3.2.5)

4.1 Communication Protocols and Virtual Observatory Standards

The ultimate goal of telescope networking is to maximize the science output from a distributed collection of instruments by coordinating their activity using a set of intelligent software agents. As the number of robotic telescopes and networks supported by individual projects is growing, so does the need for greater connectivity between heterogeneous collections of instruments and for better communication standards. There is a clear trend to move away from proprietary data formats and protocols relying on binary data representations and toward adopting portable text-based specifications that can take advantage of modern technologies for representing content such as XML (Extensible Markup Language). Connectivity is normally achieved by implementing astronomy-specific protocols for establishing connections and exchanging data that utilize existing lower level mechanism such as SOAP (Simple Object Access Protocol),

RSS (Real Simple Syndication), or XMPP (eXtensible Messaging and Presence Protocol) on top of basic TCP/IP sockets (e.g., Klotz 2010).

There are two major types of protocols involved. The first type is primarily for requesting new observations and sending commands to remote hardware and the second serves as the medium for sharing scientific information about interesting events. Remote Telescope Markup Language⁶ (RTML) created in 1999 by UC Berkeley's Hands-On Universe project is an XML dialect that supports a transparent use of remote and/or robotic telescopes (Hessman 2006b; Hessman et al. 2006). The language handles instrument capability assessments, scheduling requests, and status collecting functions. RTML provides a high level of telescope interoperability and is not intended for driving individual hardware components such as motors, filter wheels, sensors, etc. It has been adopted as the main hardware centric communication protocol by RoboNet and integrated with e-STAR (Sect. 4.5). RTML is now available in commercial astronomy software like Bob Denny's ACP2 internet telescope package. Other means of using XML to dynamically drive hardware are NASA's AIML (Astronomical Instrument Markup Language) developed for the SOFIA Observatory and INDI (Instrument Neutral Distributed Interface).

The grand daddy of rapid event dissemination systems is GCN (GRB Coordinates Network) created in 1997 (Barthelmy 2008). Its main purpose is to enable rapid follow-up of gamma-ray bursts and other high-energy transients detected by satellites such as Swift and Fermi. There are several types of GCN alerts including e-mail notifications and instant messages to cell phones and pagers. A GCN notice is a packet of 160 bytes structured as 40 long integers delivered over a live TCP/IP socket and intended for triggering hardware. Typical latency between the time a trigger is received by the GCN system and the time it is delivered to recipients is only a fraction of a second. The rules of decoding these binary transmissions are quite complicated and depend on the type of event (encoded in the first 4 bytes). Each node is configured with a mask of packet types to be transmitted combined with some simple criteria such as the size of the error box and the maximum delay between the event and the localization alert. Certain packet types are used to distribute pointing updates for instruments providing GCN triggers so that the follow-up telescopes on the ground can track the field of view of gamma-ray monitors. The main GCN process acts as a client and establishes persistent connections to all nodes configured to receive alerts treating them as servers. While this arrangement may provide more control at the GCN gateway, it is very inconvenient in modern environments with security firewalls. Regardless of its limitations, GCN has been tremendously successful and directly contributed to many groundbreaking discoveries.

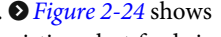
Proliferation of time-domain surveys created a need for an event messaging system designed to serve a wide range of projects. In 2006, the International Virtual Observatory Alliance (IVOA) officially adopted VOEvent as a standard format to report time-sensitive observations of astronomical objects. VOEvent messages written in XML are readable to both machines and humans and capture metadata related to the observations, as well as the inferences derived from those observations. A valid VOEvent transmission contains the following elements:

- <Who> – the author and the publisher of this information
- <How> – description of the instrumental setup
- <What> – actual measurements and other numerical data
- <Why> – scientific interpretation of the event
- <WhereWhen> – time and location of the event

⁶<http://www.astro.physik.uni-goettingen.de/hessman>

The VOEvent specification draws on other Virtual Observatory standards like Space-Time Coordinate (STC), VOTable, and VOTimeSeries. Reports of follow-up observations can be linked to messages that triggered them by cross-referencing VOEvents. Provisions were made for updates and retractions. VOEventLib is a reference Python implementation and parser for VOEvent.

4.2 Real-Time Event Brokers

The job of event brokers is to provide a platform for sending, receiving, and archiving events. VOEventNet at Caltech was the first system of this kind. Its successor Skyalert achieved a considerable level of sophistication and is scalable to handle many publishers and subscribers exchanging events at a high rate (Williams et al. 2009). The system uses modern transport protocols to communicate with both humans (RSS, Twitter, XMPP) and machines (SOAP, XML, TCP/IP sockets). It is built on top of the Python web mega-framework Django with MySQL database as the storage back end. One of the main challenges that must be addressed by event brokers is filtering a diverse flow of events so that subscribers only receive information they are interested in. Subscribers configure a set of rules to decide which events will be sent to them (or their robots). Those triggering rules are expressed in Python syntax and may utilize any information in the database, including cross-correlations between multiple data sources (e.g., equivalent to “Catalina transients without SDSS detections”). Users can annotate events with new information emerging from subsequent follow-up observations.  [Figure 2-24](#) shows several screenshots of the online interface to Skyalert. A standard set of existing alert feeds is accessible by general public. Registered Skyalert users can also create customized alert feeds. Setting up a new event stream takes some assistance from the support team. Publishers arrange to make remote event submissions using either a Java client distributed by Skyalert or a custom implementation of the protocol. For iPhone enthusiasts, there is an application called Transient Events that communicates with the Skyalert database and allows users to receive notifications of transient events shortly after they are discovered. As of 2011, over a dozen event streams are connected to Skyalert including optical transient searches (Catalina, “Pi of the Sky”), microlensing surveys (OGLE, MOA), and UV/X-ray/ γ -ray satellite observatories (GALEX, Swift).

4.3 Automated Transient Detection and Classification

Although astronomers are typically early adopters of hardware technology (especially new detectors), they are surprisingly slow to introduce more sophisticated statistical methods in their work. Perhaps this is why machine learning is so underutilized in astronomy compared to other disciplines such as biology or market research. But the situation is changing and time-domain surveys are leading the way. Accurate real-time transient classification can be achieved by assimilating on the fly the required context information: multicolor time-resolved photometry, host galaxy and other environment data, broadband spectral properties, and features across the wavelength range (Mahabal et al. 2008b). Neural nets, decision trees, and kernel methods such as support vector machines have been shown to perform well in low-level event recognition and quality assurance where there is little need to handle missing features. Bayesian belief networks are particularly useful for high-level classification work dominated by the semantics of the scientific interpretation of the data. At this stage, the system must fuse all available data from a very heterogeneous set of databases. This means that each data vector contains

The screenshot displays the SkyAlert.org website interface. At the top, it features the SkyAlert.org logo and navigation links such as 'Browse Event Streams', 'Browse SkyAlert Feeds', and 'my Feeds and Alerts'. The main content area is divided into several sections:

- Recent Events:** A section with a header 'Recent Events' and a sub-header 'In the picture below, time is measured with "right now" at the right. Ages of recent events -- the last 200 received -- are shown by stream. Click on an event to bring up a new window with detailed portfolio.' Below this is a dark-themed visualization of event streams for various telescopes: MOA, CRTS3, CRTS, CSS_NEO, Fermi, CRTS2, and CBAT. A time axis at the bottom shows 'Time since now (2011/04/25 6:16 PDT)'. Navigation buttons for 'Browse Event Streams', 'Browse SkyAlert Feeds', and 'my Feeds and Alerts' are present.
- About SkyAlert:** A text block explaining that SkyAlert collects and distributes astronomical events in near-real time. It describes how events are grouped into streams and how users can browse streams and events themselves. It lists features like 'Alerts' which decide which events are interesting, 'Atom feeds' for interesting events, and the ability to view recent events as a table or with a 'WorldWide Telescope'.
- SkyAlert News:** A section with a header 'SkyAlert News' and a sub-header 'Event IVORN: no.ihvo.cathexivovventes/tdofoboup463088 from stream: 3D55'. It includes a 'Table of Parameters' with columns for 'group', 'name', 'unit', and 'value'. The table lists various parameters like RA, Dec, positionError, etc. To the right of the table is a 'Portfolio' section for a specific event, 'CRTS (Catalina) 019050', showing a grid of images and a 'Finding Chart'.
- Transient CSS110414:075414+313216:** A section with a header 'Transient CSS110414:075414+313216' and a sub-header 'Mag: 14.3'. It includes a 'Light Map' showing a scatter plot of magnitude (Mag) versus Date (MJD - 53249). The plot shows two distinct groups of points, one in red and one in blue, representing different measurements. Below the plot are links for 'Discovery Data', 'Current Applications', 'Find and plot discovery CSS images', 'CSS110414', 'Images from other surveys', and '2012 follow-up'. A 'Pre-discovery' section mentions 'Catalina Sky Survey could image (transient location marked with IV is towards the top and II is to the left)'. To the right of the text are several small images of the transient, some with labels like 'e1', 'e4', 'e16', 'e17', and 'e18'.

Fig. 2-24

Screenshots of the SkyAlert web pages showing a transient from the Catalina Real-Time Transient Survey. Users are able to display light curves, finding charts, discovery images, and cross-identification data from external databases

potentially hundreds of features, but most of the time the actual values will be missing for most of the features. Bayesian nets are a natural choice for these high-dimensional but sparsely populated data spaces. An example application to astronomical time-domain data is the classification engine developed for the Catalina survey (Mahabal et al. 2008a). The RAPTOR/Thinking Telescopes project at Los Alamos is developing a real-time classification system for explosive transients based on the same approach.

4.4 Follow-Up Optimization

Much of the early experience with robotic systems performing autonomous follow-up comes from GCN. In case of GRB response, the rate of alerts is low and localizations are generated by

an expensive asset such as an orbiting satellite, and then distributed to less expensive instruments. The potential scientific payoff is very high and typically far outweighs the risk of wasting telescope time. In this regime, we can afford the simplest strategy “all kids go after the ball.” Even modest instruments like “Pi of the Sky” and RAPTOR survey arrays monitoring large areas of sky down to 12–15 mag are capable of generating terabytes of images per week and finding several candidate objects for follow-up at any given moment. The emerging Peta-scale surveys such as Pan-STARRS and LSST bring a new set of challenges. LSST is expected to deliver ~50,000 variability alerts per night and will quickly saturate the world supply of follow-up instruments. This bottleneck is particularly severe for spectroscopy that provides the key to unraveling the physics of the event. Humans simply lack the time, attention span, memory, and reaction speed to handle this volume of real-time data or short time scale characteristic for explosive phenomena. Increasingly, the expert observer is only involved at the highest level, and most processing takes place without human interaction. Taking humans out of the loop is not an easy task, but a number of promising results have been obtained using information-theoretic approach (Mahabal et al. 2008a). Limited data leads to degenerate event classifications with ambiguous probability distributions that must be resolved with follow-up observations taking into account instrument capability, cost, and availability, e.g., due to local weather. A successful follow-up strategy can be obtained by choosing observations that maximize the decrease in conditional entropy of the posterior probability distribution for object class. This is equivalent to maximizing the conditional mutual information of the next observation about the desired classification given previously collected data. In other words, observations with flat probability distributions contain little information about the nature of the object, while those with bumpy histograms take us closer to the final answer. A first-generation system based on this algorithm has been deployed on the data streams from the Catalina survey and PTF, clearly demonstrating the utility of this approach.

4.5 Global Telescope Networks and Deployable Computing

Globally distributed systems of a few relatively loosely connected telescopes are becoming increasingly common (Hessman 2006a). Las Cumbres Observatory Global Telescope Inc. (LCOGT) is a privately funded, nonprofit organization with the ambitious goal of building and fielding a network of about 50 telescopes located at 6–7 sites around the world (Hidas et al. 2008). The network will have a complete coverage of the night sky to “keep you in the dark” and will be used for education and time-domain astrophysics. The 2-m Faulkes telescopes built by Telescope Technologies Ltd. (TTL), one on Haleakala, Hawaii, and one at Siding Springs, Australia, are now part of the network and will be equipped with new instruments for visible/NIR imaging and spectroscopy. Currently, the LCOGT team is developing a fleet of 1-m telescopes (3 per site) for optical imaging. Each site will also host five or more 0.4-m telescopes based on Meade RCX400 f/8 tubes dedicated primarily to education. LCOGT will be primarily a follow-up facility with the remaining time used for variability studies.

Tightly integrated networks of instruments that work together without human intervention to accomplish a specific science goal are much less common. Good examples are the first generation of RAPTOR telescopes and “Pi of the Sky.” Both systems run transient detection algorithms based on coincidence checks between redundant telescopes observing the same field of view and use the parallax information to reject nearby objects. The distinguishing feature of the RAPTOR/Thinking Telescopes network at Los Alamos is its self follow-up capability

enabled by the TALONS (Telescope Alert Operations Network System) infrastructure engineered as a central bidirectional hub (White et al. 2004). A very different approach was taken by eSTAR (eScience Telescopes for Astronomical Research), a peer-to-peer network developed and operated by the University of Exeter, UK (Allan et al. 2006b). eSTAR is a completely decentralized collection of relatively independent software agents, each performing a well-defined task such as resource discovery, delivering observation requests, reducing and analyzing data, or detecting the presence of objects of some type. TALONS is built around the push model for data transport to minimize the latency of response, while eSTAR implements both push and pull transfers. The eSTAR system was used to coordinate a network of four telescopes for deep follow-up: UKIRT, Liverpool Telescopes, and the two Faulkes telescopes. In 2005, eSTAR and TALONS joined forces, opened a gateway between the two proprietary networks, and thus created the first astronomical meta-network (White et al. 2006). The community effort to federate existing telescope networks and individual instruments into a global meta-network is now led by the Virtual Observatory and Skyalert.

Modern observational science is becoming very data intensive. Astronomy is on the leading edge of the data deluge that threatens to overwhelm our computing and storage capabilities. Due to limitations of network bandwidth, data-intensive and time-critical processing is moving closer to data-generating instruments in the field, creating serious constraints on power and physical footprint. There is a renewed interest in high-density storage media based on non-volatile memory technologies such as Flash and phase-change memory. Computing industry is developing radically new architectures designed around tightly integrated processing/storage boards. While these solutions will provide the horse power necessary to handle the data at observing sites, they create even more need for intelligent software that integrates a set of telescopes, databases, and compute clusters into coherent systems for scientific exploration.

5 Science with Robotic and Survey Telescopes

Robotic telescopes found regular use in at least a dozen different types of projects. A breakup by primary science of 152 projects is listed in [Table 2-4](#). The list has been compiled by Frederic Hessman and is available online.⁷ General purpose remote observing and educational services account for almost a third of robotic telescope usage. About a quarter of instruments on the list are devoted to optical observations of cosmic explosions, mainly GRBs and supernovae. Another large group is devoted to exoplanet searches and stellar variability surveys.

5.1 Cosmic Explosions

The launch of the Swift satellite in 2004 marks the beginning of a new era in gamma-ray burst research. Follow-up observations with fast-slewing optical telescopes on the ground deliver well-sampled multicolor light curves reaching into the critical first minutes of the explosion and provide accurate localizations before the object is too faint to grant a spectrum. As a result, the number of bursts with spectroscopic redshifts increased dramatically to nearly a hundred (Xiao and Schaefer 2011). Autonomous robotic telescopes played an important role in discoveries of

⁷<http://www.uni-sw.gwdg.de/~hessman/MONET/links.lst>

■ Table 2-4

Robotic telescopes of the world by project science

Primary purpose	# telescopes	Fraction (%)
Gamma-ray bursts	31	20.4
Service observations	26	17.1
Education	20	13.2
Exoplanet searches	18	11.8
Photometric monitoring	14	9.2
All-sky surveys	12	7.9
Supernovae search	10	6.6
Asteroids	8	5.3
Spectroscopy	4	2.6
Astrometry	4	2.6
AGN, Quasars	4	2.6
(Micro)lensing	1	0.7

prompt optical/IR emission simultaneous with gamma rays (Vestrand et al. 2005), color evolution of the early afterglow (Perley et al. 2008), and delayed explosive activity in GRBs (Woźniak et al. 2006). GRB 080319B, the so-called naked-eye burst is the most luminous object known to humanity. It was found automatically by the “Pi of the Sky” system and was also observed by TORTORA and RAPTOR wide-field monitors. These modest instruments provided key data on the event including deep limits on optical precursors and a high fidelity light curve starting before the high-energy trigger (Racusin et al. 2008; Woźniak et al. 2009). Using polarization detectors on robotic telescopes, we are beginning to probe the strength and topology of magnetic fields responsible for the synchrotron emission in GRB afterglows. The Liverpool Telescope made the first two detections and found a high degree of polarization at the level of 10% (Steele et al. 2009).

For many years, the world supply of supernova discoveries was dominated by galaxy-targeted surveys with robotized instruments like KAIT. Automated nontargeted searches acquired new momentum after the discovery of accelerating expansion of the universe and dark energy. In 2010 alone, there were 538 supernova discoveries, most of them by automated searches like Catalina Real-Time Transient Search and Palomar Transient Factory (Gal-Yam and Mazzali 2011). There is a renewed interest in novae and the existence of events that populate the luminosity gap between novae and supernovae (Kasliwal et al. 2011). An automated supernova search with ROTSE-III telescopes turned up several anomalously bright supernovae that could be powered by theoretically predicted pair instability (Quimby et al. 2011). More members of this new class have been found by PTF. Learning about the diversity of cosmic explosions and dark energy is one of the major drivers for large-scale synoptic surveys such as Pan-STARRS and LSST.

5.2 Solar System Science

The current moving object searches are focused on detection of potentially hazardous asteroids (PHAs) (Stokes et al. 2002). Over the past decade, the number of known near-Earth asteroids

increased tenfold to nearly 10,000, with $\sim 10\%$ characterized as large (1 km or larger). Objects passing through the Earth-Moon system are now routinely found. The number of all moving objects discovered by leading searches (LINEAR, Catalina, Spacewatch) is measured in hundreds of thousands. Scientific benefits include a better characterization of asteroid populations as tracers of the formation and dynamical history of the solar system, new studies of comets, and identifying potential targets for flight projects (Bottke et al. 2002). The discovery volume for objects below a few 100 m in size is very small. Guarding against these types of hazards will require continuous searches. Robotic telescopes are a perfect match for this work. Deep synoptic sky surveys discover moving objects serendipitously. LSST will build a comprehensive inventory of the solar system including 80% of PHAs larger than 140 m, $\sim 20,000$ trans-Neptunian objects (TNOs) down to 100 km in diameter and millions of main-belt asteroids (MBAs) (Jones et al. 2009).

5.3 Extrasolar Planets

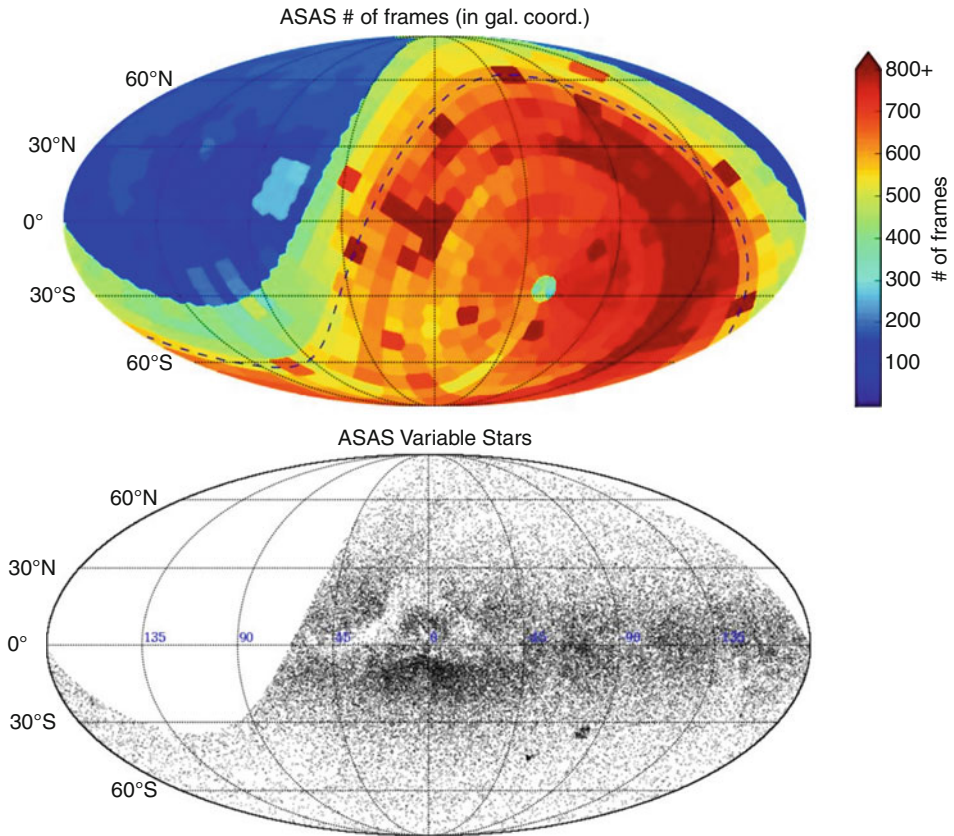
Wide-field robotic telescopes of modest size are well suited for high-cadence photometric monitoring of stellar fields in search of tiny dips in light due to transits of planets in front of their parent stars. Such systems require special data processing pipelines performing sophisticated detrending of systematic effects to suppress the noise down to a few millimagnitudes. So far, nearly 100 transiting planetary systems have been discovered with ground-based instruments of this type, mostly WASP and HAT (The Extrasolar Planets Encyclopaedia⁸). Due to the eclipsing geometry, the majority of planets found this way are hot jupiters on tight orbits with periods of a few days. Unlike other exoplanet detection methods, the transit method reveals the size of the planet. The disadvantage of this technique is low probability of transits in randomly oriented systems and high rate of false positives due to low mass stellar companions and field crowding. High preponderance of hot jupiters so close to their parent stars is forcing the theorists to rethink the models of planet formation and evolution (Mordasini et al. 2011).

Microlensing surveys found a few extrasolar planets around stars that gravitationally lensed other stars in the background (Bennett 2008). The main advantage of this method is great sensitivity to very small Earth-sized planets and potentially habitable systems. However, very low intrinsic probability of microlensing and poor prospects for observing parent stars in most events limit the usefulness of this method.

5.4 Variable Stars

Large homogeneous samples of high-quality light curves collected by modern CCD-based surveys are a gold mine of information on stellar variability. One of the many by-products of microlensing surveys (OGLE, MACHO, EROS) are extensive catalogs of variable stars in the Magellanic Clouds and in the vicinity of the galactic bulge, reaching down to $V \sim 21$ mag and collectively listing over hundred thousand objects (e.g., Soszyński et al. 2011; Alcock et al. 2004). Small robotic telescopes will soon provide a complete census of bright variables of major types down to about 15th magnitude (● Fig. 2-25). ASAS alone released a database of 50,000 light curves, tripling the number of known variable stars in this magnitude range (Pojmanski 2002).

⁸<http://exoplanet.eu>



■ Fig. 2-25

Statistics of the All-Sky Automated Survey (ASAS) in Galactic coordinates (as of early 2010): the locations of ASAS variable stars, mostly brighter than $V = 12.5$ mag (*bottom*) can be compared to the number of images available for a given line of sight (*top*). The total duration of the survey is more than a decade

Many examples of rare objects are now available for analysis such as RR Lyrae and Cepheids displaying odd combinations of pulsational modes (Poleski et al. 2010), Nova and dwarf Nova eruptions (Pojmański 2009), or R Coronae Borealis stars that can suddenly disappear in a cloud of dust produced in their atmospheres (Tisserand et al. 2011). Entirely new classes of variability have been discovered, e.g., OSARGs (OGLE Small Amplitude Red Giants; Wray et al. 2004). This wealth of new material on stellar pulsation and eclipsing systems has given us a better understanding of the structure and evolution of both single stars and binaries. A good example is the recent discovery of a classical Cepheid in a well-detached, double-lined eclipsing binary that confirmed the agreement between the dynamical mass and the pulsation mass with 1% accuracy, providing strong support for pulsation theory (Pietrzynski et al. 2010). New samples of standard candles and detached binaries are enabling improvements to distance scale in the Local Group.

6 Conclusions and Future Outlook

Sky surveys continue to play a central role in astronomy. The current trend toward automation and autonomous observing will likely accelerate over the next decade. The ultimate goal is to provide a continuous record of the entire sky with ever increasing sensitivity. Arrays of small robotic telescopes are progressively eliminating gaps of coverage that still exist for objects as bright as 12–15 mag and are opening for exploration the domain of very short time scales, from minutes down to video rates at 10–20 Hz. The availability of inexpensive detectors and optical components is enabling a dynamic growth in this area. Rapid follow-up instruments are evolving toward larger apertures and more sophisticated instrumentation that can deliver high fidelity simultaneous multicolor imaging, spectroscopy, and polarization measurements. While traditional target of opportunity programs will continue to play a vital role, in the following years, we will witness a global proliferation of dedicated rapid follow-up networks of 2-m class imagers and low-resolution spectrographs. Sky monitoring projects of the future must integrate state-of-the-art information technology such as computer vision, machine learning, and networking of the autonomous hardware and software components (Borne 2008). Emerging standards for communication (RTML, VOEvent, Skyalert) are gradually integrated into working systems.

The next generation of wide-field surveys is positioned to revolutionize the study of astrophysical transients by linking heterogeneous surveys with a wide array of follow-up instruments and by rapid dissemination of transient alerts using a variety of mechanisms on the Internet. The main challenges ahead of massive time-domain surveys are timely recognition of interesting events in the torrent of imaging data and maximizing the utility of the follow-up observations. Accurate event classification can be achieved by assimilating on the fly the required context information from external catalogs and real-time alert feeds from other instruments, both in space and on the ground: multicolor time-resolved photometry, galactic latitude, host galaxy information, broad-band spectral properties from radio to high-energy γ -rays, etc. But in order to apply this approach to extremely data-intensive projects of the future, a fundamental change is required in the way astronomy interacts with information technology. Major investments are required in the development of hierarchical, distributed decision engines capable of understanding and refining information such as partially degenerate event classifications and time-sensitive constraints on follow-up assets. The need to delegate increasingly complex tasks to machines is stimulating concentrated efforts on a number of fronts:

- Fault-tolerant system architectures and network topologies that maximize the usability
- Better classification and anomaly detection algorithms for time-variable astronomical objects
- Improved standards for real-time communication between heterogeneous hardware and software agents
- Logical and physical spatiotemporal database design with support for data fusion
- New ways of evaluating and reporting the most important science alerts to humans

Experience shows that a major leap in capability of telescopes sooner or later leads to exciting and totally unanticipated results. As far as we can tell, technology development will maintain its dazzling pace for some time. Robotic and survey telescopes are certainly at the forefront of discovery today and will remain there in the foreseeable future.

References

- Akerlof, C., et al. 1999, *Nature*, 398, 400
- Akerlof, C. W., et al. 2003, *PASP*, 115, 132
- Allcock, C., et al. 2004, *AJ*, 127, 334
- Allan, A., Bloom, J. S., & Seaman, R. 2006a, *Astron Nachr*, 327, 741
- Allan, A., Naylor, T., & Saunders, E. S. 2006b, *Astron Nachr*, 327, 767
- Allan, A., et al. 2006c, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 6274 (Bellingham, WA: SPIE)
- Bakos, G. Á., Lázár, J., Papp, I., Sári, P., & Green, E. M. 2002, *PASP*, 114, 974
- Bakos, G., et al. 2009, in IAU Symposium, Vol. 253, IAU Symposium (Cambridge: Cambridge University Press), 354–357
- Barthelmy, S. 2008, *Astron Nachr*, 329, 340
- Bennett, D. P. 2008, Detection of Extrasolar Planets by Gravitational Microlensing, ed. J. Mason, (Berlin: Springer), 47–88
- Beskin, G., et al. 2010, *Adv Astron*, 2010, Article ID 171569
- Bloom, J. S., Starr, D. L., Blake, C. H., Skrutskie, M. F., & Falco, E. E. 2006, in ASP Conf. Ser. 351, *Astronomical Data Analysis Software and Systems XV*, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique (San Francisco, CA: ASP), 751
- Bloom, J. S., et al. 2009, *ArXiv e-print* 0905.1965
- Borne, K. D. 2008, *Astron Nachr*, 329, 255
- Bottke, W. F., Jr, Cellino, A., Paolicchi, P., & Binzel, R. P. 2002, in *Asteroids III*, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. Binzel (Tucson: University of Arizona press), 3
- Bowman, M. K., Ford, M. J., Lett, R. D. J., McKay, D. J., Mücke-Herzberg, D., & Norbury, M. A. 2002, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 4848, ed. H. Lewis (Bellingham, WA: SPIE), 137–147
- Carusi, A., Gehrels, T., Helin, E. F., Marsden, B. G., Russell, K. S., Shoemaker, C. S., Shoemaker, E. M., & Steel, D. I. 1994, in *Hazards Due to Comets and Asteroids*, ed. T. Gehrels, M. S. Matthews, & A. M. Schumann (Tucson: University of Arizona press), 127
- Castro-Tirado, A. J. 2010, *Adv Astron*, 2010, Article ID 824731
- Christian, D. J., et al. 2006, *Astron Nachr*, 327, 800
- Dalton, G. B., et al. 2006, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 6269 (Bellingham, WA: SPIE)
- Djorgovski, S. G., et al. 2008, *Astron Nachr*, 329, 263
- Eisenstein, D. J., et al. 2011, *AJ*, 142, 72
- Emerson, J. P., & Sutherland, W. J. 2010, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 7733 (Bellingham, WA: SPIE)
- Filippenko, A. V., Li, W. D., Treffers, R. R., & Modjaz, M. 2001, in ASP Conf. Ser. 246, *IAU Colloq. 183: Small Telescope Astronomy on Global Scales*, ed. B. Paczynski, W.-P. Chen, & C. Lemme (San Francisco, CA: ASP), 121
- Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, *AJ*, 111, 1748
- Gal-Yam, A., & Mazzali, P. 2011, *ArXiv e-print* 1103.5165
- Gehrels, T. 1991, *Space Sci Rev*, 58, 347
- Gilmore, K., et al. 2008, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 7014 (Bellingham, WA: SPIE)
- Greiner, J., et al. 2008, *PASP*, 120, 405
- Guidorzi, C., et al. 2006, *PASP*, 118, 288
- Gunn, J. E., et al. 1998, *AJ*, 116, 3040
- Gunn, J. E., et al. 2006, *AJ*, 131, 2332
- Hearnshaw, J. B., et al. 2006, in *The 9th Asian-Pacific Regional IAU Meeting*, ed. W. Sutan-tyo, P. W. Premadi, P. Mahasena, T. Hidayat, & S. Mineshige (Bandung, Indonesia: Department of Astronomy and the Bosscha Observatory, Institut Teknologi Bandung Press), 272
- Hessman, F. V. 2006a, *Astron Nachr*, 327, 763
- Hessman, F. V. 2006b, *Astron Nachr*, 327, 751
- Hessman, F. V., Tuparev, G., & Allan, A. 2006, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 6270 (Bellingham, WA: SPIE)
- Hidas, M. G., Hawkins, E., Walker, Z., Brown, T. M., & Rosing, W. E. 2008, *Astron Nachr*, 329, 269
- Hill, G. J., MacQueen, P. J., Palunas, P., Barnes, S. I., & Shetrone, M. D. 2008, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 7014 (Bellingham, WA: SPIE)
- Hodapp, K. W., et al. 2004, *Astron Nachr*, 325, 636
- Ivezic, Z., Tyson, J. A., Allsman, R., Andrew, J., Angel, R., & for the LSST Collaboration. 2008, *ArXiv e-print* 0805.2366
- Jones, R. L., et al. 2009, *Earth Moon and Planets*, 105, 101
- Kaiser, N., et al. 2010, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, SPIE Conf. Ser. 7733 (Bellingham, WA: SPIE)

- Kasliwal, M. M., Cenko, S. B., Kulkarni, S. R., Ofek, E. O., Quimby, R., & Rau, A. 2011, *ApJ*, 735, 94
- Klotz, A. 2010, *Adv Astron*, 2010, Article ID 496765
- Kubanek, P. 2010a, *ArXiv e-print* 1002.0108
- Kubanek, P. 2010b, *Adv Astron*, Article ID 902484
- Larson, S., Spahr, T., Brownlee, J., Hergenrother, C., & McNaught, R. 1999, in *Advanced Maui Optical and Space Surveillance Technologies Conference (Maui, Hawaii: The Maui Economic Development Board)*, 182–186
- Law, N. M., et al. 2009, *PASP*, 121, 1395
- Lewis, I. J., et al. 2002, *MNRAS*, 333, 279
- Lipunov, V., et al. 2010, *Adv Astron*, 2010, Article ID 349171
- LSST Science Collaborations et al. 2009, *ArXiv e-print* 0912.0201
- Mahabal, A., et al. 2008a, *Astron Nachr*, 329, 288
- Mahabal, A., et al. 2008b, in *Am. Inst. of Phys. Conf. Ser. 1082, American Institute of Physics Conference Series*, ed. C. A. L. Bailer-Jones (New York: American Institute of Physics), 287–293
- Malek, K., et al. 2010, *Adv Astron*, Article ID 194946
- Molinari, E., et al. 2010, *Adv Astron*, 2010, Article ID 253675
- Mordasini, C., Alibert, Y., Klahr, H., & Benz, W. 2011, in *EPJ Web of Conf. 11, Detection and Dynamics of Transiting Exoplanets*, St. Michel l'Observatoire, ed. F. Bouchy, R. Díaz, & C. Moutou (Paris: EDP Sciences), id.04001, 11, 4001
- Naylor, T., Allan, A., & Steele, I. A. 2006, *Astron Nachr*, 327, 741
- Onaka, P., Tonry, J. L., Isani, S., Lee, A., Uyeshiro, R., Rae, C., Robertson, L., & Ching, G. 2008, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 7014 (Bellingham, WA: SPIE)
- Paczyński, B. 2000, *PASP*, 112, 1281
- Paczyński, B. 2001, in *Mining the Sky*, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann (Berlin: Springer), 481
- Pérez-Ramírez, D., Nemiroff, R. J., & Rafert, J. B. 2004, *Astron Nachr*, 325, 568
- Perley, D. A., et al. 2008, *ApJ*, 672, 449
- Pietrzyński, G., Thompson, I. B., Gieren, W., Graczyk, D., Udalski, G. B. A., Soszynski, I., Minniti, D., & Pilecki, B. 2010, *Nature*, 468, 542
- Plokhotnichenko, V., Beskin, G., Karpov, S., Bondar, S., de-Boer, V., Lioubetski, A., & Badjin, D. 2010, *Adv Astron*, Article ID 109681
- Pojmanski, G. 1997, *AcA*, 47, 467
- Pojmanski, G. 2002, *AcA*, 52, 397
- Pojmański, G. 2009, in *ASP Conf. Ser. 403, The Variable Universe: A Celebration of Bohdan Paczynski*, ed. K. Z. Stanek (San Francisco, CA: ASP), 52
- Poleski, R., Soszyński, I., Udalski, A., Szymański, M. K., Kubiak, M., Pietrzyński, G., Wyrzykowski, Ł., & Ulaczyk, K. 2010, *AcA*, 60, 179
- Pollacco, D. L., et al. 2006, *PASP*, 118, 1407
- Quimby, R., et al. 2011, *Nature*, 474, 487
- Racusin, J. L., et al. 2008, *Nature*, 455, 183
- Richmond, M., Treffers, R. R., & Filippenko, A. V. 1993, *PASP*, 105, 1164
- Seppala, L. G. 2002, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 4836, ed. J. A. Tyson & S. Wolff (Bellingham, WA: SPIE), 111–118
- Skrutskie, M. F., et al. 2006, *AJ*, 131, 1163
- Smith, R. J., Clay, N. R., Fraser, S. N., Marchant, J. M., Moss, C. M., & Steele, I. A. 2010, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 7737 (Bellingham, WA: SPIE)
- Soszyński, I., et al. 2011, *AcA*, 61, 1
- Steele, I. A., et al. 2004, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 5489, ed. J. M. Oschmann Jr. (Bellingham, WA: SPIE), 679–692
- Steele, I. A., Mundell, C. G., Smith, R. J., Kobayashi, S., & Guidorzi, C. 2009, *Nature*, 462, 767
- Steele, I. A., Bates, S. D., Guidorzi, C., Mottram, C. J., Mundell, C. G., & Smith, R. J. 2010, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 7735 (Bellingham, WA: SPIE)
- Steinmetz, M. 2003, in *ASP Conf. Ser. 298, GAIA Spectroscopy: Science and Technology*, ed. U. Munari (San Francisco, CA: ASP), 381
- Stokes, G. H., Evans, J. B., Vighh, H. E. M., Shelly, F. C., & Pearce, E. C. 2000, *icarus*, 148, 21
- Stokes, G. H., Evans, J. B., & Larson, S. M. 2002, in *Asteroids III*, ed. W. F. Bottke, A. Cellino, P. Paolicchi, & R. P. Binzel (Tucson: University of Arizona Press), 45
- Strassmeier, K. G., et al. 2010, *Adv Astron*, Article ID 109681
- Stubbs, C. W. 2008, *Class Quantum Gravity*, 25, 184033
- Stull, C., Taylor, S. G., Wren, J., Mascareñas, D. L., & Farrar, C. 2011, *ASCE J Struct Eng*, 137
- Sumi, T. 2010, in *ASP Conf. Ser. 430, Pathways Towards Habitable Planets*, ed. V. Coudé Du Foresto, D. M. Gelino, & I. Ribas (San Francisco, CA: ASP), 225
- Tisserand, P., et al. 2011, *A&A*, 529, A118+
- Tonry, J. L., Burke, B. E., Isani, S., Onaka, P. M., & Cooper, M. J. 2008, in *Presented at the Society of Photo-Optical Instrumentation*

- Engineers (SPIE) Conference, SPIE Conf. Ser. 7021 (Bellingham, WA: SPIE)
- Tsapras, Y., et al. 2009, *Astron Nachr*, 330, 4
- Udalski, A. 2003, *AcA*, 53, 291
- Udalski, A., Kubiak, M., & Szymanski, M. 1997, *AcA*, 47, 319
- Vestrand, W. T., Theiler, J., & Woznia, P. R. 2004a, *Astron Nachr*, 325, 477
- Vestrand, W. T., et al. 2004b, in *Am. Inst. of Phys. Conf. Ser. 727, Gamma-Ray Bursts: 30 Years of Discovery*, ed. E. Fenimore & M. Galassi (Melville, NY: American Institute of Physics), 728–732
- Vestrand, W. T., et al. 2005, *Nature*, 435, 178
- Vestrand, T., et al. 2008, in *Advanced Maui Optical and Space Surveillance Technologies Conference*, ed. S. Ryan (Kihei: Maui Economic Development Board)
- White, R. R., Wren, J., Davis, H. R., Galassi, M., Starr, D., Vestrand, W. T., & Wozniak, P. 2004, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 5496, ed. H. Lewis & G. Raffi (Bellingham, WA: SPIE), 302–312
- White, R. R., Allan, A., Evans, S., Vestrand, W. T., Wren, J., & Wozniak, P. 2006, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 6274 (Bellingham, WA: SPIE)
- Williams, R. D., Djorgovski, S. G., Drake, A. J., Graham, M. J., & Mahabal, A. 2009, in *ASP Conf. Ser. 411, Astronomical Data Analysis Software and Systems XVIII*, ed. D. A. Bohlender, D. Durand, & P. Dowler (San Francisco, CA: Astronomical Society of the Pacific), 115
- Woźniak, P. R., Vestrand, W. T., Panaitescu, A. D., Wren, J. A., Davis, H. R., & White, R. R. 2009, *ApJ*, 691, 495
- Woźniak, P. R., Vestrand, W. T., Wren, J. A., White, R. R., Evans, S. M., & Casperson, D. 2006, *ApJL*, 642, L99
- Wray, J. J., Eyer, L., & Paczyński, B. 2004, *MNRAS*, 349, 1059
- Wren, J., Vestrand, W. T., Wozniak, P., & Davis, H. 2010, in *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, SPIE Conf. Ser. 7737 (Bellingham, WA: SPIE)
- Xiao, L., & Schaefer, B. E. 2011, *ApJ*, 731, 103
- York, D. G., et al. 2000, *AJ*, 120, 1579

3 Segmented Mirror Telescopes

*Jerry Nelson*¹ · *Terry Mast*¹ · *Gary Chanan*²

¹UC Observatories/Lick Observatory, Santa Cruz, CA, USA

²University of California, Irvine, CA, USA

1	<i>Introduction</i>	101
2	<i>History of Segmented-Mirror Telescopes</i>	103
3	<i>Segmentation Geometries</i>	104
4	<i>Segment Surface Asphericity</i>	106
5	<i>Segment Polishing</i>	108
6	<i>Segment Support</i>	108
7	<i>Diffraction Effects</i>	109
8	<i>Infrared Properties</i>	111
9	<i>Active Control System</i>	112
9.1	Introduction	112
9.2	Actuators	112
9.3	Edge Sensors	112
9.4	Construction of the Active Control Matrix for Keck-Type Sensors	113
9.5	Construction of the Active Control Matrix for Vertical Sensors	115
9.5.1	Singular Value Decomposition	116
9.5.2	Error Propagation	117
9.5.3	Surface Errors from SVD (for Diffraction-Limited Observing)	119
9.5.4	Tip-Tilt Errors from SVD (for Seeing-Limited Observing)	120
9.6	Focus Mode	120
10	<i>Optical Alignment</i>	121
10.1	Tip-Tilt Alignment of Segments	121
10.2	Phasing	121
10.2.1	Shack-Hartmann Phasing	122
10.2.2	Other Phasing Techniques	125
10.3	Warping Harnesses	126
10.4	Alignment of the Secondary Mirror	127

11	<i>Other Segmented-Mirror Telescopes</i>	128
11.1	Gran Telescopio Canarias	128
11.2	Hobby-Eberly Telescope and Southern African Large Telescope	128
11.3	Large Area Multi-Object Spectrographic Telescope	129
11.4	James Webb Space Telescope	130
12	<i>Giant Segmented-Mirror Telescopes</i>	130
12.1	GMT	131
12.2	TMT	131
12.3	E-ELT	133
	<i>Acknowledgments</i>	133
	<i>References</i>	134

Abstract: Constructing the primary mirror of a telescope out of segments, rather than from a monolithic piece of glass, can drastically reduce the mass of the mirror and its material costs, thereby making possible the construction of optical/infrared telescopes with very large diameters. However, segmentation also introduces a host of complications, involving the fabrication of off-axis optics, novel diffraction effects, active segment control systems, in situ aberration correction, and the optical alignment of large numbers of degrees of freedom. Progress in these latter areas over the last 25 years has led to the successful development of the two Keck telescopes, as well as several other segmented telescopes in the 10-m class. Giant segmented telescopes of similar design, but with mirror diameters of 30–40 m, are now in the planning stages, with first light expected around the end of the decade. Segmentation has also made possible the 6.5-m James Webb Space Telescope, which is currently under construction. In this work, the technical issues associated with segmentation are discussed and reviewed in detail. Particular attention is paid to the properties of arrays of hexagonal segments (the segmentation pattern of choice for these telescopes), including their diffraction patterns and algorithms for their active control. Optical alignment issues are also discussed.

1 Introduction

The motivation for building larger ground-based telescopes has been clear for a very long time: for a telescope of fixed (seeing-limited) resolution, the observing time necessary to reach a given signal-to-noise ratio varies as $1/D^2$, where D is the diameter of the primary mirror. However, with the development of adaptive optics (AO), the motivation for building ever larger telescopes is even more powerful. In the first place, with AO, the resolution is no longer set by the atmosphere, independent of the telescope diameter, but rather is given by the diffraction limit λ/D , where λ is the wavelength, so that the angular information content now increases dramatically as the telescope gets larger. In addition, at least for background-limited adaptive optics observations, the benefit of a large telescope grows as D^4 , not D^2 ; not only is there the collecting area advantage, according to which the signal grows as D^2 , but there is less background “behind” the image, so that the noise is reduced by this same factor.

Building a giant telescope from a single monolithic mirror presents many difficulties. These difficulties typically grow rapidly with increasing mirror size and make building monolithic mirrors with diameters of 10 m or more highly impractical. The key issues, briefly stated, are:

- Mirror blank material is expensive and availability may be limited.
- The financial risk associated with breakage increases rapidly with the mirror diameter. The probability of breakage may also increase.
- Passive support of the mirror will result in large optical deflections.
- Larger mirrors are subject to larger deformations and hence large optical aberrations from thermal changes.
- The vacuum chamber for mirror coatings becomes very large and expensive.
- Tool costs for all parts (fabrication and handling) are large.
- Shipping a very large mirror to the observatory site may be impractical.

An obvious solution to these and other problems is to construct the primary mirror from smaller segments, rather than building a single large mirror. Consider problems related to the

mass or thickness of the mirror. The gravitational deflection suffered by an optic increases as D^4 . This means that a monolith must have N times the thickness of a segmented mirror consisting of N equal segments covering the same area. Thus, the 36 segments of each Keck telescope are 7.5 cm thick; an equivalent monolith would be 2.7 m thick! Although many important problems are greatly reduced by building the primary from small segments, there are a number of issues, concerns, and problems that arise with segmentation, and these must be understood and dealt with before one can confidently proceed with building a large telescope having a segmented primary.

The various issues and complications associated with segments can generally be grouped into four categories:

- Because of the large number of segments (36–91 in the current segmented telescopes, 500–1,000 in the extremely large telescopes of the future), the telescope will have a large increase in the number of parts and a proportional increase in complexity.
- Segments are difficult to polish because they are off-axis sections of the parent figure of revolution and are therefore not locally axisymmetric.
- Segments require both active position control and optical alignment. Here and throughout this review, these concepts are distinguished as follows. Optical alignment places the mirror segments in the proper relative positions with respect to one another to within optical tolerances; active position control keeps these relative positions fixed (i.e., it “freezes” the mirror).
- Segment edges and the associated intersegment gaps add to both diffraction and thermal background effects.

The first item, increased complexity, is a general one and is only addressed in a general way in this review. To some extent, the increase in complexity is offset by the advantage, not shared by monolithic systems, that small prototype subsystems, consisting, for example, of a few (full-sized) segments, together with their mechanical supports and associated control electronics, can be built relatively cheaply and used for thorough testing and optimization. In this way, one can engineer components that are highly reliable, so that the probability of failure of critical components can be kept relatively low.

The complexity problem can be further mitigated by providing sufficient spares and by designing the system so that one is able to replace defective components quickly and easily, especially in those cases where the failure of even a single component of a given type (e.g., a segment position actuator) would have serious consequences. It may also be possible to design some aspects of the system so that the failure of a few components of a given type will have little or no effect on the performance of the overall system. Thus, the active control systems of the Keck telescopes are sufficiently robust that the failure of several edge sensors can be tolerated (as long as they can be identified and removed from the control loop). Indeed, neither Keck telescope has ever operated with its full complement of 168 edge sensors all working at the same time. Note that for extremely large telescopes even the temporary loss of an actuator may have little impact on many kinds of science.

The last three items in the above list are specific issues that are discussed in some detail below. A brief history of segmented-mirror telescopes is presented in [▶ Sect. 2](#). General segmentation issues are discussed in [▶ Sect. 3](#), followed by segment asphericity in [▶ Sect. 4](#), segment polishing in [▶ Sect. 5](#), and segment support in [▶ Sect. 6](#). Diffraction and thermal effects are treated in [▶ Sects. 7](#) and [▶ 8](#), respectively. Active control of segments is discussed in [▶ Sect. 9](#) and optical alignment in [▶ Sect. 10](#). Little or no time is spent on those aspects of large (or extremely large) telescopes that are not specifically related to segmentation – for

example, telescope drives or dome shutters. However, it is worth noting that adapting the conventional designs of such items to the scale of a 10-m telescope (or larger) can often be a nontrivial problem in itself.

The technical issues listed in the previous paragraph are discussed below in the specific context of the Keck telescopes but with an indication as to how these may be generalized, as, for example, in the discussion of control matrices; alternative approaches to such topics as segment phasing are also described. The review concludes with descriptions of the other segmented-mirror telescopes currently in operation in [Sect. 11](#), as well as those that are under development in [Sect. 12](#).

2 History of Segmented-Mirror Telescopes

The first recorded use of segmented optics appears to have been by Archimedes, who in 212 BC used an array of mirrors to focus the sun's rays on the ships of the attacking Roman navy in order to defend Syracuse (although claims that the ships burst into flames as a result may have been exaggerated (Papadogiannis et al. 2009)). More recently, Horn d'Arturo in Italy made a 1.5-m mirror out of 61 hexagonal segments in 1932. However, it was only used vertically and was not actively controlled (Horn D'Arturo 1955). In the 1970s, Pierre Connes in France made a 4.6-m segmented-mirror telescope for infrared astronomy (Chevallard et al. 1977). It was fully steerable and actively controlled, but there were problems with image quality, and the telescope was never put into operation.

A variation on the theme of a segmented-mirror telescope was the Multiple Mirror Telescope (MMT) (Beckers et al. 1982); it consisted of six 1.8-m telescopes in a common mount. The MMT was not a true segmented-mirror telescope and no longer exists in its original multiple mirror configuration, but its advantages and disadvantages compared to a true segmented-mirror telescope are nevertheless worthy of some discussion (see [Sect. 3](#)).

The Keck Observatory began as a conceptual design study for a 10-m segmented-mirror telescope (Nelson et al. 1985) in the late 1970s. This project was formally begun in 1984, and full science operations began in 1993. The telescope was quite successful, and as a result, funds were acquired to build a second Keck telescope, 75 m from the first. The close proximity was designed to allow the two Keck telescopes to be used for interferometry as well as for individual telescope observing. Keck 2 was completed in 1996 and began science observations in that year. Each of the telescopes has its own suite of scientific instruments and an adaptive optics system. Interferometric science observations began in 2003. The success of the Keck telescopes led to a Spanish project, the Gran Telescopio Canarias, (GTC) (Rodríguez Espinosa et al. 1999), a 10-m telescope similar to Keck on La Palma in the Canary Islands (see [Sect. 11.1](#)).

The Keck telescopes have hyperbolic primary mirrors and this leads to the requirement of polishing off-axis mirror segments. However, this is not the only possible approach to segmentation. Telescopes with spherical primaries can also be designed; these have identical, although not identically spaced, segments, and must contend with a large amount of spherical aberration. A telescope of this latter design, the Hobby-Eberly Telescope, was completed in the late 1990s at the McDonald Observatory in Texas (Barnes et al. 2000). The HET has 91 spherically polished primary mirror segments and is effectively a 9-m telescope. A telescope of very similar design, the Southern African Large Telescope (SALT) (Buckley et al. 2004) saw first light in 2005. The HET and SALT are described in further detail in [Sect. 11.2](#).

Segmented-mirror telescopes as large as 50–100 m have been proposed (Andersen et al. 2004; Dierickx et al. 2002), but the two largest such telescopes currently well into the design

stage are the Thirty Meter Telescope (TMT) (Nelson and Sanders 2006), a collaboration of the United States, Canada, and several international partners, and the 42 m European Extremely Large Telescope (E-ELT) (Gilmuzzi and Spyromilio 2008), a project of the European Southern Observatory. The Giant Magellan Telescope (circumscribed diameter 24.5 m), although of markedly different design than the other telescopes described in this review, is nevertheless considered a giant segmented-mirror telescope and is discussed briefly in [Sect. 12.1](#). The TMT and E-ELT are described further in [Sects. 12.2](#) and [12.3](#).

3 Segmentation Geometries

There are many ways one can imagine dividing up a primary mirror into smaller optical elements, including:

- Independent telescope arrays
- Independent telescopes on a common mount
- Random subapertures as part of a common primary
- Annular segmentation of a common primary
- Hexagonal segmentation of a common primary

Independent arrays of telescopes have been considered for decades but have generally not been successful, except for radio telescopes, which are aided by the fact that the individual telescope signals can be amplified and combined while preserving phase information. This is not practical in the optical; thus, there are significant inefficiencies associated with coherently combining the light from an array of optical telescopes. Instrumentation for an array of telescopes has also been a cause of difficulty. Perhaps the best known successful array has been the VLT with four 8-m telescopes, each with its own suite of science instruments, and the capacity to combine all telescopes together for interferometric measurements. (The VLT can also be operated as four individual telescopes.)

An alternative to a segmented primary is to put a modest number of independent telescopes into a common mount. The Multiple Mirror Telescope (MMT) (Beckers et al. 1982), consisting of six 1.8-m telescopes, including six separate secondaries, configured in this way (equivalent to a 4.5-m aperture), operated on Mount Hopkins in Arizona, from 1979 to 1998. Although the multiple mirror approach can result in large, well-corrected fields of view, it has many disadvantages compared to a true segmented-mirror telescope. These include the less compact design, which necessitates a larger dome and slit, the difficulty of phasing noncontiguous mirrors, the need for an expensive and complicated beam combiner to bring the light to a common focus, and a generally more complicated diffraction pattern from the sparse aperture. The original six-mirror MMT is no longer in use, having been replaced by a single 6.5-m mirror of honeycomb design (West et al. 1997) in 2000. The proposed Giant Magellan Telescope (Johns 2008) resembles the original MMT in that it has seven circular primary mirror “segments,” each 8 m in diameter, (it includes an on-axis mirror, which the original MMT did not have), but it will have a single secondary so that a beam combiner is not needed.

Special application telescopes have been built with very sparse arrays in order to sample the resolution space with the minimum number of mirrors. Systems like this may give greater angular resolution (longer baseline) but have less sensitivity due to the smaller collecting area. Dense segmentation is generally favored when one is attempting functional replication of the optical properties of a single monolithic mirror. These kinds of mirrors have the best central

concentration of light, leading to the smallest instruments and best signal-to-noise ratio, and, for a given collecting area, are generally the most compact and thus the least expensive.

Dense segmentation can be achieved with radial petals (Burgarella et al. 2002), annular patterns, hexagonal patterns, and many others. A regular pattern can only be made of regular polygons (e.g., hexagons) if the mirror is flat, since a curved surface can only be tessellated with nonregular polygons. However, the segments may appear regular when seen in projection. Each of these patterns has its advantages and disadvantages, specific to the fabrication techniques planned. For example, annular rings in general have large numbers of identical segments, so that replication techniques may be utilized and the number of necessary spares is reduced. Hexagons make the best use of the material typically produced in round boules of glass and are easier to polish because the corners are less severe than are four-sided polygons.

In general, for a telescope with N hexagonal segments, there are $N/6$ different segment shapes. Having six segments of each shape has not been enough to motivate the use of replication techniques, but at least the number of required spares is reduced by a significant factor. Hexagonal segments also tend to be easier to support against gravity and to attach to position actuators (three per segment) because of their symmetry. Overall, the advantages of hexagonal segments have made this the segmentation pattern of choice for most of the current and planned segmented-mirror telescopes.

Hexagonal segments are naturally arrayed in rings. The Keck telescopes consist of three rings with a total of 36 segments; the central segment, which would be blocked by the secondary mirror, is omitted. For a telescope of n rings, with the central segment omitted, the total number of segments is

$$N_{\text{seg}} = 3n^2 + 3n \quad (3.1)$$

and the total number of intersegment edges is

$$N_{\text{edge}} = 9n^2 + 3n - 6 \quad (3.2)$$

There are three actuators per segment and two sensors per intersegment edge. As the number of segments becomes large, the number of sensors approaches six per segment (two per actuator), since the 12 sensors around a segment are shared with its nearest neighbors. Edge effects, larger for smaller values of n , reduce this ratio somewhat. (Note that, with the central segment missing, a single ring of six segments ($n = 1$) would have fewer sensors (12) than actuators (18) and thus would not provide a stable configuration for testing a partially completed telescope. However, two rings (72 sensors and 54 actuators) is a stable configuration.) For the extremely large telescopes of the future, with very large numbers of segments, it is advantageous to omit some segments in the outer rings in order to produce a more nearly circular primary.

A significant issue in the segmentation of a telescope mirror is the determination of the optimal segment size. This involves a complicated trade-off: smaller segments are easier and less expensive to fabricate and to support, but there need to be more of them, so alignment and control becomes more difficult. The Keck design study chose a segment side length of 0.9 m, so that 36 segments were required to fill the 10-m aperture; the thickness was 75 mm. In retrospect, for these parameters, the control and alignment of the segments was in some sense easier than the fabrication. As a result, the giant segmented telescopes of the future will have somewhat smaller segments, with a hexagon side length of about 0.7 m.

4 Segment Surface Asphericity

Two-mirror telescopes are the most common optical design for ground-based telescopes. These systems require a parabolic or hyperbolic primary mirror. As mentioned above, it is possible to build a segmented-mirror telescope with a primary mirror that is spherical so that the segments are easier to fabricate; however, several additional mirrors are then needed to correct the resultant spherical aberration, and the associated light loss and additional alignment complexity make this configuration less commonly used. In this review, it is assumed that a nonspherical primary is desired and the resulting requirements on the segment figures are discussed.

The primary mirror is typically a figure of revolution, but since it is not spherical, pieces of the primary will not look identical and will not be figures of revolution about their local centers. This basic fact introduces significant complexity for segmented mirrors. The polishing of off-axis segments, which are not figures of revolution, is generally much more difficult than the polishing of spheres. Second, off-axis optics must be carefully aligned in all six rigid body degrees of freedom, and the larger the segment asphericity (deviation from a sphere), the tighter the alignment tolerances become.

This section describes in mathematical detail the surfaces of these segments. The general equation of a conic can be written as

$$r^2 - 2kx + (1 + K)z^2 = 0 \quad (3.3)$$

where r is the (global) radial coordinate, k is the radius of curvature, and K is the conic constant. It is useful to expand this in powers of r as

$$z(r) = \frac{r^2}{2k} + (K + 1)\frac{r^4}{8k^3} + (K + 1)^2\frac{r^6}{16k^5} + (K + 1)^3\frac{5r^8}{128k^7} + \dots \quad (3.4)$$

Expressed in the local segment coordinate system, the symmetry of the equation seen in the global coordinate system is lost, and one will see azimuthal variations. It is useful to express the equation for the segment surface in its local coordinate system:

$$z(\rho, \theta) = \sum_{n,m} \alpha_{nm} \rho^n \cos m\theta + \sum_{n,m} \beta_{nm} \rho^n \sin m\theta \quad (3.5)$$

where ρ, θ are the local coordinate system polar coordinates. A suitable rotation of the local coordinate system will cause the β_{nm} to vanish, and analyticity requires that $n, m \geq 0$, $n \geq m$, and $n - m$ must be divisible by 2. The expansion has been made (Nelson and Temple-Raston 1982) and yields for the expansion coefficients α_{mn} :

$$\alpha_{20} = \frac{a^2}{k} \left[\frac{2 - K\epsilon^2}{4(1 - K\epsilon^2)^{3/2}} \right] \quad (3.6)$$

$$\alpha_{22} = \frac{a^2}{k} \left[\frac{K\epsilon^2}{4(1 - K\epsilon^2)^{3/2}} \right] \quad (3.7)$$

$$\alpha_{31} = \frac{a^3}{k^2} \left[\frac{K\epsilon[1 - (K + 1)\epsilon^2]^{1/2}(4 - K\epsilon^2)}{8(1 - K\epsilon^2)^3} \right] \quad (3.8)$$

$$\alpha_{33} = \frac{a^3}{k^2} \left[\frac{K^2\epsilon^3[1 - (K + 1)\epsilon^2]^{1/2}}{8(1 - K\epsilon^2)^3} \right] \quad (3.9)$$

$$\alpha_{40} = \frac{a^4}{k^3} \left[\frac{8(1 + K) - 24K\epsilon^2 + 3K^2(1 - 3K)\epsilon^4 - K^3(2 - K)\epsilon^6}{64(1 - K\epsilon^2)^{9/2}} \right] \quad (3.10)$$

■ Table 3-1

Expansion coefficients for Keck outermost segment shapes

Coefficient	Value (μm)	Name
α_{20}	11,376	Focus
α_{22}	-101.1	Astigmatism
α_{31}	-38.1	Coma
α_{33}	0.17	Trefoil
α_{40}	0.09	Spherical aberration

where a is the segment radius, $\epsilon = R/k$, and R is the off-axis distance of the segment center. It is useful to expand each of the above equations as a power series in ϵ :

$$\alpha_{20} = \frac{a^2}{2k} + \frac{Ka^2\epsilon^2}{2k} + \frac{9K^2a^2\epsilon^4}{16k} + \dots \quad (3.11)$$

$$\alpha_{22} = \frac{Ka^2\epsilon^2}{4k} + \frac{3Ka^2\epsilon^4}{8k} + \dots \quad (3.12)$$

$$\alpha_{31} = \frac{Ka^3\epsilon}{2k^2} + \frac{(9K-2)Ka^3\epsilon^3}{8k^2} + \dots \quad (3.13)$$

$$\alpha_{33} = \frac{K^2a^3\epsilon^3}{8k^2} + \dots \quad (3.14)$$

$$\alpha_{40} = \frac{(1+K)a^4}{8k^3} + \dots \quad (3.15)$$

As an explicit example consider the outermost segment ($R = 4.68$ m) for the Keck telescopes, which has $a = 0.9$ m, $k = 35$ m, $K = -1.003683$, and $D = 10.95$ m. The segment figures are given in terms of the expansion coefficients in [Table 3-1](#). Note that the expansion here is in terms of the functions $\rho^m \cos n\theta$. These are not the same as Zernike polynomials, but for the lowest orders, they differ only in terms of normalization and in meaningless piston, tip, and tilt (and in defocus for α_{40}). (They are therefore referred to by the same names.) The former functions are used here for simplicity, for continuity with the literature, and because it is convenient in the present context that the coefficients α_{nm} give the maximum deviation of the surface directly. Elsewhere in this review, aberrations (as opposed to nominal surface shapes) are expressed in terms of Zernike polynomials. For the latter, the numbering and normalization conventions of Noll (1976) are followed. In the latter convention, the Zernike coefficient gives the rms over the circumscribed circular surface, not the maximum deviation.

The focus term simply represents the nominally spherical surface. The higher terms quantify the departures from sphericity. As the size of the segments is reduced (for a fixed overall primary mirror area), the higher order terms quickly become smaller, making the segments easier to polish, but more difficult to control, since there are more of them. For the 1.8-m-diameter Keck segments, inspection of [Table 3-1](#) shows that both the astigmatism and coma terms are quite large compared to the ~ 0.5 μm wavelength of visible light and must therefore be polished out correctly to a fraction of a percent of their nominal values. This represents a unique challenge for segmented optics and requires special polishing techniques.

5 Segment Polishing

There are two major areas of concern in polishing segments. The first is the challenge of polishing aspheres. The second is the issue of polishing the surface properly all the way out to the edge of the segment.

Polishing aspheres is difficult because polishing only works well when the polishing tool fits the glass surface to within typical distances of order $1\ \mu\text{m}$. Polishing tools move in a random motion to produce the desired smoothness, which means that the tool must be spherical in shape, and thus the contact area is limited by the asphericity. There are three main approaches to polishing. The first, the one used for the Keck segments, is stressed mirror polishing (Lubliner and Nelson 1980; Nelson et al. 1980) where the mirror is deformed so that the desired shape is mapped into a sphere. Large tools can then be used to polish a spherical surface, which will relax into the desired shape when the stresses are released. The second approach, stressed lap polishing, is to deform the tool dynamically so that it always fits the mirror as it moves around the surface. The third method is to use suitably small tools so that the tool fit is adequate and make raster scans of the mirror with this small tool.

In addition to these approaches, there are two methods currently in use that do not require a good fit between the tool and the part. These are ion beam figuring and magnetic rheological figuring (MRF). In ion beam figuring, a beam of argon or another rare gas ion is used to remove material from the optical surface, analogous to sandblasting, but at the atomic level (Braunecker et al. 2008). The technique has the advantages of being noncontact, non-iterative, and quite accurate, but it must be done in a vacuum chamber, and the rate of material removal is relatively slow. Ion figuring was used as the final production step in making the Keck segments. The MRF technique exploits the fact that a magneto-rheological fluid greatly increases its viscosity in the presence of a magnetic field, so that its ability to transmit a force can be accurately controlled and rapidly adjusted. Like ion beam figuring, MRF is accurate but relatively slow.

Edge effects must be considered carefully because a significant fraction of the overall area of the primary mirror lies relatively close to at least one intersegment edge. Again, several approaches have been considered. The approach used for the Keck segments was to polish the mirrors as rounds and, when the main polishing was complete, to cut the rounds into the desired hexagonal segments. This introduces no local effects but introduces some global deformations associated with the release of internal stresses. For the Keck segments, the latter deformations were removed by ion figuring. A second approach is to add small “shelves” of material to the edges in order to support the polishing tool when it is near the edge. It is the polishing near the edge that typically rolls the edge in an optically objectionable fashion. The third approach is to polish near the edges with smaller and smaller tools to control the size of the rolled edge. A variant on the first method is to polish spheres with a so-called planetary polisher. In this case, the polishing tool is much larger than the part, and the part is placed face down on the polishing tool in order to produce roll-free edges.

6 Segment Support

Supporting the primary mirror segments is a nontrivial problem in mechanical engineering. If a Keck segment were supported on the three actuator points alone, the maximum deflection under gravity due to the weight of the segment itself would be about $1.7\ \mu\text{m}$ when the telescope is pointed at the zenith (Nelson et al. 1985). (A formal study of the deflection of thin plates

on point supports has been made by Nelson et al. (1982)). In order to reduce the deflections by the required two orders of magnitude, the axial force of each actuator is distributed over a 12-point whiffletree, so that the segment is supported on a total of 36 points. This reduces the gravitational deflections to 8 nm. To satisfy the high-bandwidth requirements of the active control system, the whiffletrees must be very stiff axially, but they also need to be very flexible radially. The Keck whiffletree design accomplishes this by means of aluminum I-beams with flex pivots and flex rods at each of the 36 points. The flexures also minimize stress on the mirror due to thermal expansion of the whiffletrees.

By attaching springs that extend from the whiffletree to the mirror cell, forces can be applied to the segment in order to warp its surface. These so-called warping harnesses are used at Keck to provide in situ correction to the segment figures. The ability to make warping harness adjustments allows one to correct for the aberrations that segments may acquire as the result of release of internal stresses when the originally circular segments are cut into hexagons, and also allows one to relax the polishing tolerances on the segments, with a resulting cost savings. Adjustment of the Keck warping harnesses is carried out manually immediately after a segment is installed after aluminization. The measurements required for warping harness adjustments at Keck are described in [▶ Sect. 10.3](#) below.

Each Keck segment is supported radially by a thin diaphragm. As the telescope moves from zenith to horizon, the load gradually transfers from the whiffletrees to the diaphragm. The mechanical characteristics of the diaphragm are opposite to those of the whiffletrees: the diaphragm is radially stiff and axially flexible. A thick displacement-limiting disk beneath the diaphragm provides a hard stop that prevents the segment from moving far enough to damage the edge sensors.

7 Diffraction Effects

Segment edges will introduce diffractive effects in the image in addition to the diffractive effects caused by the overall aperture itself. Of course, it is desirable to minimize the size of segment gaps, but finite physical and optical gaps are both necessary. Physical (air) gaps provide clearance so that segments do not touch each other during installation and removal. Optical (non-reflective) gaps in the form of bevels are necessary to avoid chipping of edges. Both the physical and optical gaps are typically a few millimeters wide. The Keck segments have 2-mm bevels on the edges and a 3-mm air gap between segments, for a 7-mm total nonoptical strip between segments. The design for the TMT calls for gaps about half this size. By contrast, for a telescope made up of identical segments with spherical surfaces, the gaps are set by geometry, since, as previously noted, it is impossible to tessellate a curved surface with regular polygons. These latter gaps tend to be larger than those mandated by physical and optical considerations, but the concern in this section is primarily with the smaller, Keck-type gaps.

For circular apertures, the simplest of mirrors, the diffraction-limited image is an Airy pattern, and for large angular distances ω the intensity pattern falls as ω^{-3} , and is azimuthally symmetric. Polygonal mirror segments (such as those of Keck) will concentrate the diffracted energy into lines perpendicular to the segment edges, thus producing a diffraction pattern that is brighter or darker in some places than that of a circular aperture.

The amount of energy in the diffraction pattern and the angular scale of this pattern are set by the size of the segment and the size of the intersegment gaps. In general, diffracted energy is spread out over an angular scale of order λ/ℓ where λ is the wavelength and ℓ is the appropriate

linear scale. This may be the segment gap, the segment diameter, the full diameter of the primary mirror, etc. Furthermore, the fraction of the total intensity that is diffracted out to the angular scale characteristic of the segment gaps will simply be equal to the fraction of the area of the primary mirror that is covered by the gaps (including the associated edge bevels).

The Keck gaps (physical and optical combined) cover about 0.7% of the area of the primary mirror; thus, there is an added diffraction pattern that has about 0.7% of the flux in the central image. At $1\ \mu\text{m}$, this energy will be diffracted into angular scales of about 30 arcsec. The diffracted energy is small and is spread over a very large region, making its local effects even less significant. For comparison, the structural spiders that typically support the secondary mirror generally block about 1% of the light going to the primary, and since the support widths are typically 1–4 cm, this diffracted energy is larger than that of segment edges and more centrally concentrated by a factor of several. Both of these factors tend to make diffraction from secondary mirror supports more problematic than diffraction from intersegment gaps.

The diffraction pattern for a regular hexagon is proportional to the absolute value squared of the Fourier transform of the hexagon aperture function H . This function is defined by the condition $H(x, y) = 1$ when the aperture plane coordinates x and y fall within the aperture and $H(x, y) = 0$ otherwise. Let u and v be the corresponding image plane coordinates (in units of spatial frequency), $k = 2\pi/\lambda$, and s' be the side length of the hexagon, which is assumed to be centered at the origin, with two sides parallel to the x -axis. The Fourier transform is real and may be obtained analytically (Chanan and Troy 1999) as

$$\hat{H}(u, v) = \frac{2\sqrt{3}}{k^2 u} [\hat{K}_+(u, v) + \hat{K}_-(u, v)] \quad (3.16)$$

where

$$\hat{K}_\pm(u, v) = \frac{\cos(\sqrt{3}kvs'/2 \pm kus'/2) - \cos(kus')}{u \mp \sqrt{3}v} \quad (3.17)$$

For $u = v = 0$, this expression reduces as expected to the area of the hexagon: $\hat{H}(0, 0) = \frac{3\sqrt{3}}{2}s'^2$.

Now consider an array of hexagons. Let s be the side length of the hexagons which define the array (to be distinguished from the physical side length s' of the hexagonal segments), and let the vector ρ_i specify the position of the center of the i th segment in a nonoverlapping array in the aperture plane. The diffraction pattern from an array of hexagons can then be obtained directly from the pattern for a single hexagon as

$$\hat{f}(\omega) = \hat{H}(\omega) \sum_i \exp(ik\rho_i \cdot \omega) \quad (3.18)$$

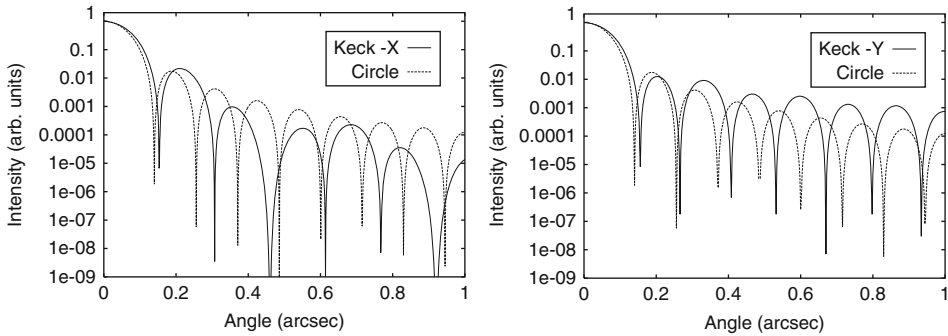
For close-packed arrays of hexagons, the coordinates of ρ_i are

$$x = \frac{3}{2}ks, \quad y = \frac{\sqrt{3}}{2}\ell s \quad (3.19)$$

where k and ℓ are integers (positive, negative, or zero) such that $k + \ell$ is even; for Keck, the 36 segment coordinates are defined by the strict inequality:

$$0 < 3k^2 + \ell^2 < 48 \quad (3.20)$$

Because there are physical (and optical) gaps between the segments, s will in general be slightly larger than s' ; the gap width is $\sqrt{3}(s - s')$. Plots of monochromatic diffraction patterns for large telescopes are often somewhat misleading because much of the fine structure (of order λ/D)



■ Fig. 3-1

Theoretical diffraction patterns for the Keck telescopes at a wavelength of $1\ \mu\text{m}$, in the direction parallel (*left panel*) and perpendicular (*right panel*) to the segment edges. The diffraction pattern corresponding to a circular aperture, shown for comparison, is quite similar

will be washed out for even relatively modest bandwidths $\Delta\lambda$. Note that this same formalism can be applied to the case of circular segments.

Theoretical monochromatic point source profiles for Keck at $1\ \mu\text{m}$ in the directions parallel and perpendicular to the segment edges are shown in [▶ Fig. 3-1](#), together with the corresponding profiles for a circular aperture of diameter 10 m. On this angular scale, the Keck and circular profiles are quite comparable. Not apparent in these figures are the effects of gaps, which show up on much larger angular scales. The issue of background surface brightness from diffraction effects in general is a complicated one and beyond the scope of this review, involving not only the effects of gaps but also the secondary mirror support structure, optical aberrations, and finite bandwidth effects. Troy and Chanan (2003) have considered this problem in the context of giant segmented-mirror telescopes.

8 Infrared Properties

At infrared wavelengths longer than $\sim 2.2\ \mu\text{m}$, the thermal emission from the environment (including the telescope and optics) becomes an important source of background. Typical environmental temperatures are of order 300 K, for which the peak emission occurs at wavelengths of about $10\ \mu\text{m}$. The short wavelength tail of the blackbody spectrum falls off exponentially, which means that the thermal background rises very rapidly from 2 to $10\ \mu\text{m}$.

All telescopes that are used in the infrared suffer to some degree from this thermal background, depending on the temperature of the telescope and its optics. In practice, telescopes with clean and freshly applied mirror coatings (such as silver) have emissivities of about 1% per surface at wavelengths beyond $1\ \mu\text{m}$. As the optics degrade with time and the accumulation of contaminants, the emissivity will grow.

The gaps between segments will generally have much higher emissivities, often close to unity. Thus, for a telescope like Keck, the segment edges will add an additional 0.7% thermal

background flux in addition to the roughly 3% associated with the three-mirror Nasmyth configuration. When one includes the typical effects of dirt and aging and the backgrounds from the atmosphere and from any additional warm optics in the beam train (such as windows or warm adaptive optics mirrors), the added background from segment edges is a real but nonetheless rather small effect.

9 Active Control System

9.1 Introduction

At Keck and other segmented-mirror telescopes, the primary mirror segments are actively positioned in their three out-of-plane degrees of freedom by three mechanical actuators, which connect the segment to its supporting subcell (Mast and Nelson 1982; Jared et al. 1990). (Because the optical tolerances on the in-plane degrees of freedom are considerably less restrictive, these three degrees of freedom are positioned passively.) The relative displacements of adjacent mirror segments are sensed by precision electromechanical edge sensors, of which there are two per intersegment edge. The segments are actively controlled (Cohen et al. 1994) by means of a two-step process: (1) initially, the desired readings of the edge sensors are determined by external optical means; (2) subsequently, the mirror is stabilized against perturbations due to gravity and thermal effects by moving the actuators so as to maintain the sensor readings at their desired values. At Keck, the actuators are updated every 0.5 s; for the large segmented telescopes of the future, the update rate is likely to be somewhat higher.

In the following subsections, the actuators and sensors used in the Keck telescopes are discussed first, followed by a description in somewhat general terms of the construction of the control matrix that relates actuator motions to sensor readings. This in turn leads naturally to a discussion of mirror modes and error propagation by the control matrix. One particular mode, focus mode, is singled out for more detailed discussion.

9.2 Actuators

In the Keck actuators, a DC motor drives a roller screw with a 1-mm pitch, which in turn supplies axial drive to a slide. The slide drives a small piston into an oil-filled chamber. On the other side of the chamber, a large piston drives an output shaft, which is attached to the mirror. The piston areas are in the ratio 24:1, and the rotary encoders have 10,000 steps per revolution, so that the actuator step size is about 4 nm. The actuator range is 1.1 mm. Each actuator is contained within a cylindrical package 15 cm in diameter and 63 cm long, with a total mass of 11.5 kg. The power dissipation is 0.5 W per actuator, mostly due to the light source in the rotary encoder (Meng et al. 1990).

9.3 Edge Sensors

The Keck edge sensors are interlocking, with one half of the sensor attached to each neighboring mirror segment. To be precise, a drive paddle mounted on one segment fits into the sensor body on a neighboring segment, with a 4-mm clearance both above and below the

paddle. (These mechanical parts are mounted on the back of the mirror and do not obscure any of the reflective surface.) The surfaces above and below the nominal 4-mm gaps are conducting so that the gaps and conducting surfaces define two capacitors. When the two segments move relative to each other, the drive paddle moves relative to the sensor body, changing the gaps and hence the capacitance. The difference between the two capacitances is related linearly to the relative displacement of the segments. A sensor preamplifier and analog-to-digital converter measure this difference in capacitance and produce a digital output proportional to the displacement. This output is sent to the control electronics for further processing.


The sensor noise (Minor et al. 1990) is only a few nanometers. The low thermal error (less than 3 nm/K), low drift rate (less than 4 nm/week), and predictable deformation of the sensor under gravity (correctable to better than 7.5 nm rms) make the overall sensor contribution to the optical error budget quite low.

9.4 Construction of the Active Control Matrix for Keck-Type Sensors

The linear relationship between the actuators and sensors in the active control system is given by

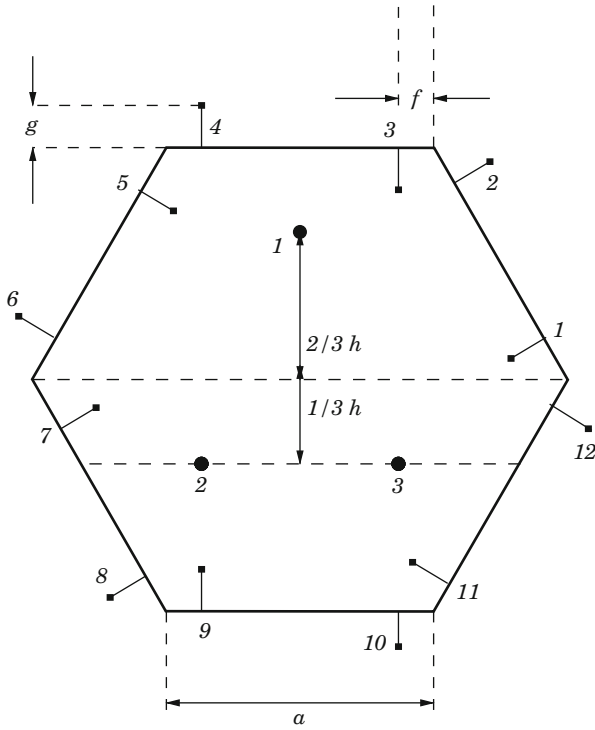
$$\mathbf{Az} = \mathbf{s} \quad (3.21)$$

where \mathbf{z} is a vector containing all of the actuator lengths (with 108 components in the case of Keck) and \mathbf{s} is a vector of all of the sensor lengths (168 components). The control matrix \mathbf{A} (168×108) is determined solely by geometry. The calculation of the precise values of the matrix elements A_{ij} is described in the following paragraph.

As noted above, the Keck sensors are horizontal, that is, the plates of the differential capacitors which make up the sensors are parallel to the segment surface. The geometrical relationships between the twelve half sensors and the segment which they monitor are defined in  Fig. 3-2; the placement of the three segment actuators is also indicated. The Keck parameters are given by $a = 900$ mm, $f = 173$ mm, $g = 55$ mm and $h = 706$ mm. The sensors sense the relative edge height, that is, the height of a segment relative to its neighbor, at the points indicated by the numbered squares in the figure. For the Keck geometry, and for most practical cases, the sensing points for sensors 7 and 12 are both above the line connecting actuators 2 and 3; the simple sign convention below is affected if this is not the case. The values of the ratios f/a , g/a , and h/a for Keck are close to optimal in the sense of minimal noise multiplication (see below), but they also reflect various practical considerations, so that the precise values do not have a fundamental significance. Practical concerns (specifically, the interchangeability of segments) may dictate that the orientation of the actuator triangle will vary from segment to segment, but for simplicity in this discussion, all actuator triangles are taken to have the same orientation. This simplification will not change the basic properties of the associated control matrix; in particular, it will have no effect at all on the error multipliers or other similar quantities derived below.

Now if actuator 1 is pistoned by an amount Δs , the segment will rotate about a line through actuators 2 and 3, so that the reading of each sensor on the segment (in height units) will change by

$$\Delta s = \frac{r \Delta z}{h} \quad (3.22)$$



■ Fig. 3-2 The geometry of the Keck active control system, showing the locations of the 3 actuators and 12 sensors halves on a typical segment


where r is the perpendicular distance from the sensor to the rotation axis and where the sign of r is positive if the sensor and the actuator are on the same side of the rotation axis and negative if they are on opposite sides.

If actuator 1 is moved by an amount Δz , then from (3.22), the corresponding edge height increment at sensor positions 1–6 will be

$$\begin{aligned}
 \Delta s_{1,1} &= \Delta z \left[\frac{1}{3}h + f \cos 30^\circ - g \sin 30^\circ \right] / h \\
 \Delta s_{2,1} &= \Delta z \left[\frac{1}{3}h + (a - f) \cos 30^\circ + g \sin 30^\circ \right] / h \\
 \Delta s_{3,1} &= \Delta z \left[\frac{1}{3}h + a \cos 30^\circ - g \right] / h \\
 \Delta s_{4,1} &= \Delta z \left[\frac{1}{3}h + a \cos 30^\circ + g \right] / h \\
 \Delta s_{5,1} &= \Delta z \left[\frac{1}{3}h + (a - f) \cos 30^\circ - g \sin 30^\circ \right] / h \\
 \Delta s_{6,1} &= \Delta z \left[\frac{1}{3}h + f \cos 30^\circ + g \sin 30^\circ \right] / h
 \end{aligned} \tag{3.23}$$

For sensor positions 7–12, $\Delta s_{j,1}$ can be obtained from $\Delta s_{j-6,1}$, but with $\sin 30^\circ$ and $\cos 30^\circ$ replaced by $-\sin 30^\circ$ and $-\cos 30^\circ$, respectively. A useful check on the signs and normalizations of the above relations is provided by various closure relations, which follow from the symmetries of the system. For example,

$$\Delta s_{1,1} + \Delta s_{5,1} + \Delta s_{9,1} = \Delta z \quad (3.24)$$

It is a straightforward matter to relate the edge height increments to the readings on the differential capacitors that constitute the edge sensors (Chanan et al. 2004). It is convenient to define the signs of the edge height differences so that all segments are similar. For actuator 1 and the sensors in  Fig. 3-2, one has

$$A_{j,1} = \pm \Delta s_{j,1} / \Delta z \quad (3.25)$$

where the sign is positive for $j = 2, 4, 6, 7, 9, 11$ and negative for the remaining values of j . The entire control matrix can readily be filled out in this way. Each actuator affects 12 sensors (or fewer for peripheral segments, since these do not have six nearest neighbors), and each sensor is affected by six actuators (in all cases). Thus, each row of the control matrix has up to 12 nonzero elements, and each column has exactly six nonzero elements, independent of the number or arrangement of segments.

9.5 Construction of the Active Control Matrix for Vertical Sensors

For the sake of simplicity and definiteness, the above discussion was presented in the context of Keck-style capacitive edge sensors. In this sensor design, the capacitor plates are horizontal and the two capacitors lie one above the other. Such sensors are interlocking, which complicates segment exchanges, and involve many parts, which makes them expensive to build. By contrast, the sensor design under consideration for the future Thirty Meter Telescope has the capacitors attached directly to the vertical sides of the segments (Mast and Nelson 2000); that is, the capacitor plates are perpendicular to the segment surface. In particular, one half of the sensor consists of a single vertical sense plate bonded or plated directly onto the side of one segment, and the other half consists of two vertical drive plates on the side of its neighbor segment directly across the intersegment gap; in effect, there are again two capacitors, one above the other. Even though there is no physical offset from the segment edge, such sensors still retain a sensitivity to dihedral angle: a change in dihedral angle will affect the two capacitor gaps differently. There is therefore an effective offset, although its geometrical interpretation is not as simple as before, and it tends to be smaller than for the Keck-style sensors. (The TMT sensor design calls for an effective offset of about 25 mm, compared to the actual offset of 55 mm for the Keck sensors.) The construction and properties of the active control matrix for vertical capacitive sensors are similar to those for the horizontal sensors described above (Chanan et al. 2004), and the same is true as well for sensors that utilize induction, not capacitance, to measure relative displacement. For primary mirrors of sufficiently small focal ratio, it may be necessary to take the curvature of the mirror into account explicitly, as opposed to using the planar approximation considered above. However, the planar approximation is adequate for Keck.

9.5.1 Singular Value Decomposition

The control equation (3.21) directly gives the sensor values corresponding to the actuator lengths. Implementing the actual control requires solving the inverse problem: what (changes in the) sensor readings are required to produce the desired (changes in the) actuator lengths? If the control matrix could be inverted, the desired solution would be obtained as

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{z} \quad (3.26)$$

However, for an overdetermined system, the matrix \mathbf{A} is not square and its inverse does not exist; for that matter, in general, an exact solution does not exist. Nevertheless, the least-squares solution can be constructed from the *pseudo-inverse* matrix, which does exist and can be constructed by straightforward means and which will still be denoted by \mathbf{A}^{-1} . A particularly useful technique for constructing the pseudo-inverse is singular value decomposition (SVD) (Golub and van Loan 1996; Press et al. 1989; Anderson et al. 1999) of the original control matrix. This technique is briefly reviewed here. In SVD, the $m \times n$ matrix \mathbf{A} (where $m \geq n$) can be written as the product of three matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T \quad (3.27)$$

where \mathbf{U} is an $m \times n$ column orthogonal matrix, \mathbf{W} is an $n \times n$ diagonal matrix whose diagonal elements w_i are positive or zero and are referred to as the singular values of the matrix \mathbf{A} , \mathbf{V} is an $n \times n$ orthonormal matrix, and the symbol T denotes transpose. The pseudo-inverse is then obtained as

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T \quad (3.28)$$

where the j th diagonal element $1/w_j$ of \mathbf{W}^{-1} is replaced by 0 in the event that $w_j = 0$. The matrix \mathbf{V} defines an essentially unique orthonormal basis set of modes of the system, such that any arbitrary configuration of the system can be expressed as a unique linear combination of these modes. In particular, V_{ij} gives the value of the i th actuator in the j th mode.

For the control matrices considered here, three of the singular values are identically equal to zero; the corresponding singular modes are the three actuator vectors corresponding to rigid body motion (global piston, tip, and tilt) of the primary mirror as a whole (since such motion has no effect on the sensor readings). If there are N segments, there are $3N$ actuators and $3N - 3$ modes of interest in the basis set.

Once the elements of the pseudo-inverse matrix are known, control of the mirror is implemented as follows: The desired sensor readings are defined when the alignment of the telescope is correct as determined by external optical means; actuator lengths are changed further only to maintain these desired sensor readings in the face of deformations due to gravity and temperature changes. The actuator changes that will maintain the desired sensor readings are calculated with the aid of the pseudo-inverse matrix elements A_{ji}^{-1} via

$$\Delta z_j = \sum_i A_{ji}^{-1} \Delta s_i \quad (3.29)$$

where the symbol Δz_i refers to the difference between the actual and desired actuator values and similarly for the corresponding differences Δs_i in the sensor readings.

In principle, the control matrix only needs to be inverted once and for all time; in practice, it has to be re-inverted every time a sensor is removed from or added to the control loop. At Keck, such changes typically happen every few months or so. By contrast, maintaining the sensor

readings at their desired values requires only a simple matrix multiply. This must be done at the frequency of the control loop, 2 Hz for Keck.

The computational power required to support the control algorithm increases rapidly with the size of the telescope. While the Keck active control system uses 168 sensors to control 108 actuators, the Thirty Meter Telescope will require 2,772 sensors to control 1,476 actuators.

9.5.2 Error Propagation

The SVD analysis can readily be extended to describe error propagation in the control system. If one were to put random uncorrelated noise equally into all sensors, then the actuators would respond proportionally as determined by the A -matrix:

$$\delta a = \alpha \delta s \quad (3.30)$$

where δs and δa are the rms values of the sensor and actuator errors, and we refer to the dimensionless parameter α as the (overall) noise multiplier. Alternatively, one could put random noise into the sensors and determine the rms amplitude $\delta \alpha_k$ for each of the above $3N - 3$ modes.

By the orthogonality of the modes, one has

$$\delta a^2 = \sum_k \delta a_k^2 = \sum_k \alpha_k^2 \delta s^2 \quad (3.31)$$

It is convenient to order the modes (the columns of the matrix \mathbf{V}) according to the magnitude of their error multipliers, from largest to smallest, or – what is the same thing – according to the size of the singular values from smallest to largest. With this ordering, it is convenient to define a residual error multiplier r_k , which includes the error multiplier of the k th mode and all higher modes:

$$r_k^2 = \sum_{j \geq k} \alpha_j^2 \quad (3.32)$$

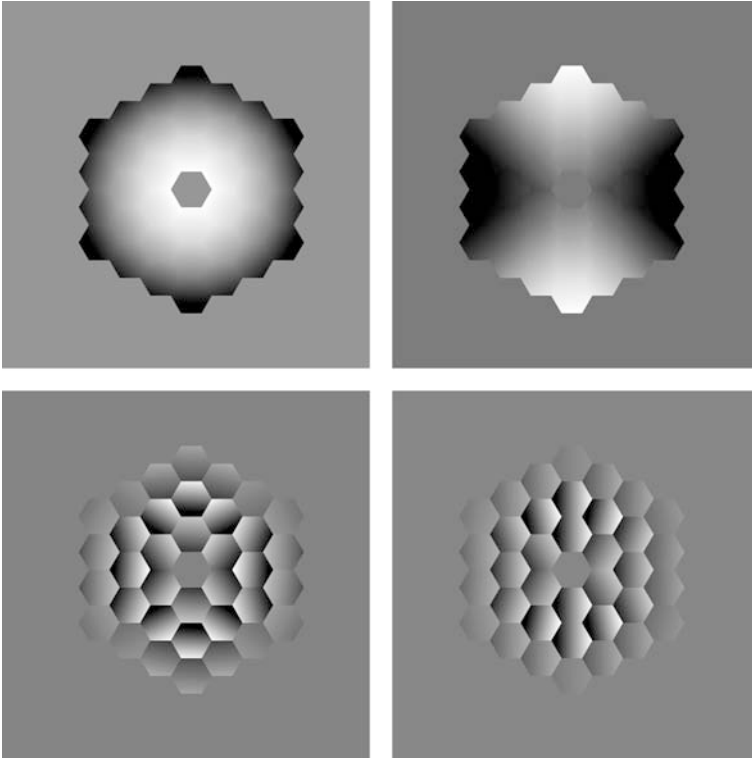
(Note that r_1 is then the same as the overall error multiplier α .) When the modes are ordered in this way, they are also more or less ordered in spatial frequency from lowest to highest. The reason for this correspondence is not hard to understand: low spatial frequency modes have small edge discontinuities that are difficult for the sensors to detect and therefore for the active control system to control; high spatial frequency modes have large edge discontinuities that are easily detected by the edge sensors. The error multiplier α_j associated with the j th mode can be shown to be

$$\alpha_j^2 = \frac{n}{w_j^2} \quad (3.33)$$

For a fixed number of segments, the individual error multipliers α_j and the overall error multiplier α are (exactly) independent of the hexagon side length a .

🔍 *Figure 3-3* shows the two Keck modes with the lowest spatial frequencies (largest error multipliers) and the two highest spatial frequency modes (smallest error multipliers). These results are typical of all sensor geometries. Inspection of the modes shows that there is a close correspondence between the lowest order modes and the Zernike polynomials.

The full range of error multipliers for the three-ring Keck telescope, a five-ring telescope similar to the Hobby-Eberly Telescope (HET), and a TMT-type (492 segment) telescope is plotted in 🔍 *Fig. 3-4*. For directness of comparison, the identical sensor geometry has been assumed

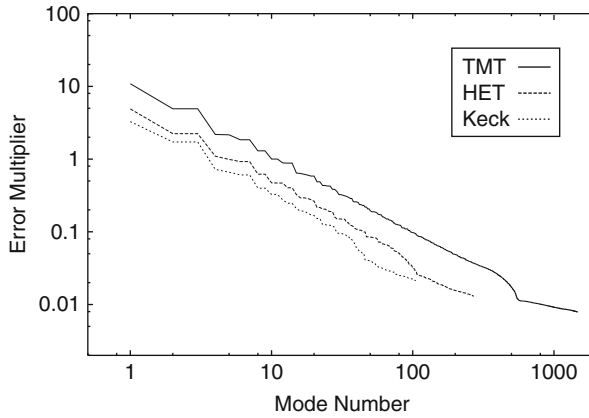


■ Fig. 3-3

Top panels: the two lowest spatial frequency modes of the Keck telescope active control system. The mode on the *upper left* is focus mode. This mode has continuous edges and the slope discontinuities are too small to be seen. The mode on the *upper right* is one of two global astigmatism modes. This mode has small edge discontinuities in addition to the slope discontinuities. The small size of the discontinuities makes these modes the most difficult to control. *Bottom panels:* the two highest spatial frequency modes. The large edge discontinuities make these modes the easiest to control

in each case. Except for focus mode, however, the multipliers are only weakly dependent on sensor geometry. The error multiplier curves scale approximately as the square root of the total number of segments; equivalently, the singular value for a mode of particular spatial frequency is, for low spatial frequencies, virtually independent of the number of segments.

Because of the sharp increase in error multiplier with decreasing mode number, one might worry that for extremely large segmented telescopes, effective mirror control will require optically determined wavefront information to supplement the electromechanical sensors. Although the issue has not yet been completely resolved for the giant telescopes currently in the design stages, it seems likely that little or no supplemental wavefront control for giant segmented telescopes will in fact be necessary (MacMartin and Chanan 2002, 2004). The argument may be summarized as follows. For diffraction-limited (adaptive optics) observing, a wavefront sensor will already be present as a part of the AO system, and it is expected that a reasonable



■ Fig. 3-4

Error multipliers for the control systems for TMT, HET, and Keck. For directness of comparison, the Keck actuator and sensor geometry has been assumed for each control system, together with the actual segment geometry for the telescope

AO system will necessarily have the dynamic range and bandwidth to correct automatically any residual low spatial frequency errors left over from the active control system, without the need for a supplementary alignment wavefront sensor. For seeing-limited observations, one will not have an AO wavefront sensor to correct the misalignments automatically, but the large aberrations from atmospheric turbulence will likely dominate those due to residual misalignments from the control system.

9.5.3 Surface Errors from SVD (for Diffraction-Limited Observing)

For diffraction-limited observing, a useful optical figure of merit for the telescope is the rms wavefront error, which is equal to twice the rms surface error. Note that here the averaging should be over the entire surface; the actuator calculation presented above averages the surface only over the discrete points corresponding to the actuator locations. However, one can show that the true rms surface error is very nearly equal to the rms actuator error.

Let z_{ij} represent the displacement of the i th actuator ($i = 1, 2, 3$) on the j th segment, let

$$p_j = (z_{1j} + z_{2j} + z_{3j})/3 \quad (3.34)$$

represent the piston error on that segment, and let \tilde{z} represent the continuous height variable over the segment surface, where $\tilde{z} = z_i$ at the actuator locations. Averaging over all segments and over an ensemble defined by a Gaussian distribution of sensor (not actuator) errors yields

$$\langle \tilde{z}^2 \rangle = \kappa \langle z^2 \rangle + (1 - \kappa) \langle p^2 \rangle \quad (3.35)$$

where

$$\kappa = \frac{15}{16} \frac{a^2}{h^2} \quad (3.36)$$

and $\langle z^2 \rangle$ denotes the average over the discrete actuator points only. If the actuators on a given segment were uncorrelated, one would have $\langle z^2 \rangle = \langle p^2 \rangle$ and thus $\langle \tilde{z}^2 \rangle = \langle z^2 \rangle$, and there would be no distinction between the rms actuator and the true rms surface. In fact, if the state of the mirror is defined by the sensors, which have a Gaussian distribution of errors, then the actuators will be weakly correlated and there will be a (small) difference between the rms surface and the rms actuator. In general, however, the difference is very small (in theory about 3% for the Keck telescopes), and there is little point in maintaining a distinction between these two quantities. The rms actuator is preferred on the grounds of simplicity.

9.5.4 Tip-Tilt Errors from SVD (for Seeing-Limited Observing)

For seeing-limited observations, one is interested not in the rms surface error (as discussed above) but rather in the rms segment tip/tilt, or the rms ray tip/tilt, where the latter includes the factor of two for doubling on reflection. The rms ray tip/tilt may be obtained directly from the SVD formalism (Chanan et al. 2004). The results may be summarized as follows.

While the error multipliers discussed above are independent of the segment size (for a fixed number of segments) and scale as the square root of the total number of segments, the tip/tilt errors scale inversely as the segment size, can vary by 50% or so depending on sensor geometry, and are virtually independent of the number of segments. The typical rms one-dimensional image blur is 2–3 milliarcsec per nm of sensor noise for a segment with $a = 0.500$ m or 1–1.5 milliarcsec per nm for a segment with $a = 1$ m.

9.6 Focus Mode

Although it is convenient to think of the edge sensors as responding to the vertical shear of one segment with respect to its neighbor, in general, the sensors will also respond to a change in the dihedral angle between adjacent segments. If all of the segment dihedral angles are changed by the same amount, or (what is essentially the same thing) if a constant is added to all of the sensor readings, this defines a defocus-like configuration of the primary mirror that is referred to as *focus mode*. In focus mode, the radius of curvature of the surface defined by the segment centers does not match the individual segment radii of curvature. Focus mode corresponds to one of the **V**-modes of the singular value decomposition, usually the mode with the smallest nonzero singular value and hence the largest error multiplier; that is, it is the least well-controlled mode. (If the dihedral angle sensitivity goes to zero, the singular value of this mode also goes to zero, and it becomes truly unobservable.) The problem may be aggravated by practical considerations. For example, if all sensors had a small but common temperature dependence, then a change in temperature could lead spontaneously to the introduction of focus mode.

However, focus mode can be corrected to first order by pistoning the secondary mirror. This will leave the wavefront with an overall residual scalloping, but since the aberration is quadratic in the coordinates, the scalloping amplitude will be smaller than the original focus mode amplitude by a factor of N , the number of segments. At Keck, this means that focus mode only has to be corrected very occasionally (typically once per month); in the interim, it can be corrected by conventional focusing of the telescope and one simply tolerates the small residual scalloping.

10 Optical Alignment

10.1 Tip-Tilt Alignment of Segments

The measurement of segment tip/tilt angles is similar to wavefront sensing in adaptive optics but with one important difference. As in adaptive optics, a Shack-Hartmann wavefront sensor is used, with either single or multiple subapertures per segment (Chanan 1988). The difference is that in adaptive optics, one is essentially trying to freeze the atmosphere, so that the exposures need to be very short (typically a few milliseconds); in segment alignment measurements, one is trying to average over local atmospheric tip/tilt errors, so that the exposures need to be quite long, typically tens of seconds.

Although the optical effects of turbulence are well understood theoretically in the short exposure limit, (Fried 1966; Noll 1976) the theoretical situation is much less clear for the long (but not infinitely long) exposures that are typical of alignment measurements. Measurements made at the VLT suggest that the residual atmospheric aberrations (up to Zernike terms of order $n \sim 4$) for long exposures are well described by the same statistical distribution of Zernike coefficients as in the theoretical short exposure limit but with an effective atmospheric coherence diameter r_0 as a single free parameter (Noethe 2002). Although the effective r_0 for these long exposures may be as much as an order of magnitude larger than the usual “instantaneous” r_0 , residual atmospheric turbulence nevertheless still limits the alignment process at the VLT. Similar results have been found at Keck, for length scales on the order of a segment. Outer scale effects, which tend to become important beyond 10 m (Ziad et al. 2004), may ameliorate this situation somewhat for the extremely large telescopes of the future.

10.2 Phasing

The most challenging degrees of freedom to align optically are those associated with segment piston errors. In order that the telescope achieve the diffraction limit corresponding to the full aperture, as opposed to the diffraction limit of the individual segments, the piston errors (or steps between segments) must be small compared to the wavelength of observation. The Strehl ratio for a segmented telescope whose N perfect segments are aligned in tip/tilt but not piston is

$$S = \frac{1 + (N - 1)e^{-\sigma^2}}{N} \quad (3.37)$$

where σ is the standard deviation of the piston errors measured in radians at the wavefront (not at the segment surface) (Chanan and Troy 1999). For a 10-m telescope with adaptive optics in the near-IR, the corresponding segment piston error requirement is less than about 30 nm for high angular resolution imaging; for the giant segmented telescopes of the future, it may be as small as 10 nm.

In principle, a segment with a piston error will produce an out-of-focus image, but this effect is extremely small: at the $f/15$ focus of the Keck telescope, the individual segment beams have a depth of focus of some 10 mm, which exceeds the dynamic range of the segment pistons by a factor of several. One is instead forced to exploit diffraction effects from misaligned intersegment edges. At Keck, this is accomplished by a physical optics generalization of the traditional geometrical optics Shack-Hartmann test. The Keck technique is described in the [Sect. 10.2.1](#) below. Recently, several other techniques have been developed, and these are described in [Sect. 10.2.2](#).

10.2.1 Shack-Hartmann Phasing

The basic idea of Shack-Hartmann phasing is to place the subapertures so that they straddle the intersegment edges of the (reimaged) primary mirror (Chanan 1988). The two halves of the subaperture are analogous to the slits in Young's two-slit interference experiment. The details of the resulting interference or diffraction pattern are sensitive to the relative piston error between the two segments. (In principle, relative tip/tilt errors between the two segments can also be detected with Shack-Hartmann phasing, but this possibility is not examined here.)

For simplicity, consider the interference pattern from a single subaperture and assume that an adaptive optics system either is not used or is located downstream from the image plane under consideration here. This means that the phasing subimages will be affected by atmospheric turbulence and this limits the phasing subaperture diameter d to less than about $r_0(\lambda)$, the atmospheric coherence diameter at the wavelength at which the phasing measurements are made. At Keck, this wavelength is about $0.9 \mu\text{m}$, and the subapertures, referred to the primary mirror, are 12 cm in diameter. Atmospheric tip/tilt errors across a subaperture (the dominant atmospheric error for $d \leq r_0$) are generally the limiting factor for geometrical optics Shack-Hartmann tests, but have little effect on phase measurements. As a result, phasing measurements do not require the long exposure times that are necessary to eliminate segment tip/tilt errors. Typical exposure times for phasing measurements at Keck are 15 s, although the question of the minimum exposure time has not been thoroughly explored.

A modest amount of analysis will clarify the basic physics of Shack-Hartmann phasing. Let $\boldsymbol{\rho} = (\rho, \theta)$ be the position vector in the subaperture plane and let $\boldsymbol{\omega} = (\omega, \phi)$ be the position vector in the image plane. (It is assumed that the components of $\boldsymbol{\rho}$ have units of length and components of $\boldsymbol{\omega}$ have units of radians.) Consider a circular subaperture of radius a straddling two segments separated by the horizontal diameter of the circle, with a physical step height δ between the two. (The corresponding wave-front step height is 2δ .) In the absence of other aberrations, the complex amplitude of the wavefront in the image plane $\hat{f}(\boldsymbol{\omega}; k\delta)$ is the Fourier transform of the complex aperture function $f(\boldsymbol{\rho}; k\delta)$:

$$f(\boldsymbol{\rho}; k\delta) = \begin{cases} \exp(+ik\delta) & \text{for } \rho \leq a; 0 \leq \theta < \pi \\ \exp(-ik\delta) & \text{for } \rho \leq a; \pi \leq \theta < 2\pi \\ 0 & \text{for } \rho > a \end{cases} \quad (3.38)$$

where $k = 2\pi/\lambda$ and the normalization is chosen such that the integral of the intensity over the image is unity. The intensity in the image plane is then (Chanan et al. 1998)

$$I(\boldsymbol{\omega}; k\delta) = [\cos k\delta \hat{f}(\boldsymbol{\omega}; 0) + \sin k\delta \hat{f}(\boldsymbol{\omega}; \pi/2)]^2 \quad (3.39)$$

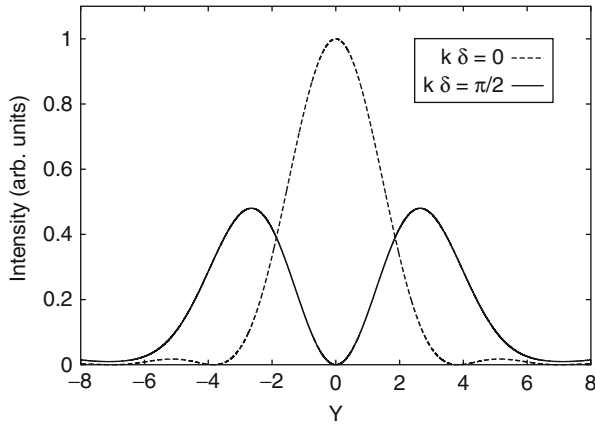
where the in-phase amplitude ($k\delta = 0$) is that corresponding to the familiar Airy disk:

$$\hat{f}(\boldsymbol{\omega}; 0) = \frac{2J_1(ka\omega)}{ka\omega} \quad (3.40)$$

and the out-of-phase amplitude ($k\delta = \pi/2$) is given by

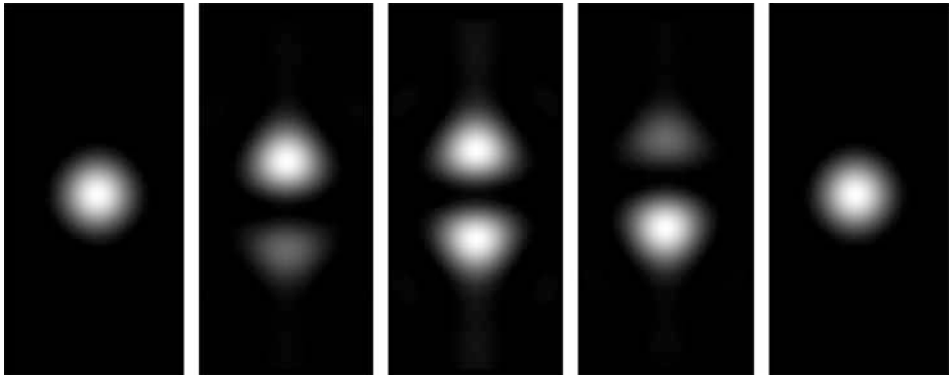
$$\hat{f}(\boldsymbol{\omega}; \pi/2) = \frac{2}{\pi} \int_0^\pi \frac{u \cos u - \sin u}{u^2} d\theta \quad (3.41)$$

where $u = ka\omega \cos(\theta - \phi)$, with both angles defined as above. This integral can be evaluated explicitly for only a few values of ϕ . For $\phi = 0$ or $\phi = \pi$, the integral vanishes; for $\phi = \pm\pi$, it reduces to $\mp H_1(ka\omega)/(ka\omega)$, where H_1 is the Struve function of order 1. However, these four



■ Fig. 3-5

Slices along the y -axis (perpendicular to the intersegment edge) of the phasing diffraction patterns for in-phase ($k\delta = 0$) and (maximally) out-of-phase segments ($k\delta = \pi/2$). In the former case, the image is the familiar Airy disk; in the latter, it splits into two equal images. The horizontal units are $k\lambda\omega$. The vertical units are arbitrary but the same for both slices



■ Fig. 3-6

Theoretical diffraction patterns from a single phasing subaperture with relative piston errors of 0, 150, 200, 250, and 400 nm. The wavelength is 800 nm and the subaperture diameter is 120 mm. The figure may be read left to right or right to left, depending on the sign convention for the piston error

cases are sufficient to show that for $k\delta = \pi/2$, the image splits into two equal subimages, with the intensity vanishing everywhere along the x -axis (see [Figure 3-5](#)).

For phase differences other than $k\delta = 0$ and $k\delta = \pm\pi/2$, the image splits into two *unequal* subimages; the ratio of the intensities of the subimages provides a measure of the phase difference or step height between the two segments. [Figure 3-6](#) shows the two-dimensional theoretical diffraction patterns for several steps between $k\delta = 0$ and $k\delta = \pi$. The middle

panel ($k\delta = \pi/2$) corresponds to a step height of $\lambda/4$, which is the maximal effect. The step height may be extracted from the diffraction pattern via a cross-correlation (Chanan et al. 2000) against a sequence of numerically generated template images. Alternative analytical extraction techniques have been discussed in the literature (Schumacher et al. 2002).

It is clear from physical considerations and from the equations that the above monochromatic phasing technique can only give answers for δ in the range 0 to $\lambda/2$ or equivalently (and more conveniently for our purposes) $-\lambda/4$ to $\lambda/4$. All piston errors of larger absolute value will be aliased into this latter interval. There are two different approaches to resolving the resulting piston ambiguity: multiple wavelength measurements and broadband measurements.

In the multiwavelength approach, one makes monochromatic measurements at several discrete wavelengths and looks for a solution that is consistent with all of the measurements. This is also referred to as the artificial wavelength method, since for the case of two wavelengths, the capture range is the same as that for a single effective wavelength λ_a :

$$\lambda_a = \frac{\lambda_1\lambda_2}{|\lambda_1 - \lambda_2|} \quad (3.42)$$

where λ_1 and λ_2 are the two individual wavelengths. For large initial piston errors, the technique can be extended to more than two wavelengths. There is an optimal way to choose the individual wavelengths that depends on the initial piston uncertainty as well as on the uncertainty in the individual measurements (Lofdahl and Eriksson 2001).

By contrast, in the broadband approach (Chanan et al. 1998), a continuous range of wavelengths is used. In this case, one is interested in the coherence of the signal. (The precise mathematical definition of coherence is not essential to the present discussion; see Chanan et al. (1998) for one possible definition.) The coherence length of the filter, which sets the wavelength bandwidth, is defined by

$$\lambda_c = \frac{\lambda_1\lambda_2}{2\Delta\lambda} \quad (3.43)$$

The diffraction pattern will then resemble that of the monochromatic case for $\delta \ll \lambda_c$. Conversely, if $\delta \gg \lambda_c$, the details of the diffraction pattern will be washed out, and the image will be incoherent; the incoherent image can be obtained theoretically by taking the monochromatic result for a nonzero edge step and averaging over all wavelengths in the bandpass or by taking the nominal wavelength and averaging over all possible phases. In broadband phasing, the primary mirror segments are stepped through a series of different configurations such that every intersegment edge changes from its nominal value by $2j\lambda_c/(n+1)$, where j takes on all integer values from $-(n-1)/2$ to $+(n-1)/2$ and n is an odd integer of moderate size; the Keck broadband phasing procedure uses $n = 11$. The overall capture range is a bit less than $\pm\lambda_c$. Practical considerations make it difficult to decrease the initial piston error uncertainty by more than about a factor of 30 in a single cycle of 11 measurements with a given filter. At Keck this means that a few cycles are needed to reduce the piston errors from their initial values of order $10\ \mu\text{m}$ to the required $30\ \text{nm}$. Each cycle involves a filter of broader bandwidth, hence shorter coherence length. The detailed characteristics of each cycle used at Keck may be found elsewhere (Chanan et al. 1998).

Although the broadband technique is more time consuming to execute than the multiwavelength technique and the extraction of the edge step from the data is not as straightforward, the former has a distinct advantage: If the initial piston error is underestimated, then the technique will fail in a well-defined and easily recognized way: most or all of the signals from the edge in question will be incoherent. However, if the initial errors or the measurement uncertainties are underestimated in the multiwavelength technique, then it will often converge to the wrong

answer, with no easily generated diagnostic. For this reason, the broadband technique has been preferred at Keck.

A variation of the broadband technique called dispersed fringe sensing (DFS) uses a grism (a transmission grating replicated onto a prism) to disperse the light along each intersegment edge, so that a continuum of wavelengths can be observed simultaneously, and there is no need to step through different configurations of the primary mirror. A prototype of this technique was successfully tested at the Keck telescope (Shi et al. 2004). The DFS technique was proposed as the coarse segment phasing technique for the James Webb Space Telescope, but ultimately a different though closely related technique, dispersed Hartmann sensing, was selected instead (see ● Sect. 11.4).

10.2.2 Other Phasing Techniques

In late 2008 and early 2009, the European Southern Observatory mounted an active phasing experiment (APE) at the Very Large Telescope in order to compare three alternative phasing techniques to the Shack-Hartmann approach describe in the preceding section (Gonte et al. 2004). The 8-m aperture of one of the VLTs was reimaged onto a 150 mm mirror with 61 actively controlled segments, so that the testing could be done under conditions of realistic atmospheric turbulence. The three techniques involve variations of curvature sensing and pyramid sensing, as well as a modified Mach-Zehnder interferometer. Unlike Shack-Hartmann sensing, which perturbs the wavefront in the pupil plane, these techniques perturb the wavefront in the focal plane (except for curvature sensing, which effects its wavefront perturbation simply by going out of focus) and relay the pupil plane to the detector. They all therefore eliminate the problem of very tight pupil registration in the Shack-Hartmann approach, although in general the determination of the precise mapping between points on the primary mirror and the corresponding points on the detector is more complicated for these techniques than for Shack-Hartmann phasing. The additional phasing techniques tested by ESO are briefly described as follows:

- *Curvature Sensing.* Roddier (1988) developed a wavefront sensing technique based on measuring the spatial variation of the intensity of defocused images; this gives a measure of the Laplacian of a continuous wavefront from which, with suitable boundary conditions, the wavefront can be reconstructed. For the discontinuous wavefronts from segmented mirrors, piston errors can be reconstructed using Fresnel diffraction theory (Chueca et al. 2008). A variation of this technique has been implemented at infrared wavelengths at Keck, with some success (Chanan et al. 1999).
- *Pyramid Sensing.* In this wavefront sensing technique, developed for use in adaptive optics, a four-faceted pyramid in the focal plane splits the beam so that four separate pupil images are formed. It is in some sense a quantitative version of the classical knife edge test but with knife edges in two orthogonal directions at the same time. The signal is a measure of the first derivative of the wavefront and thus again has good piston sensitivity (Esposito et al. 2005).
- *Mach-Zehnder Interferometer.* In this technique, a spatially filtered beam from the telescope is interfered with a phase-shifted version of itself (Montoya 2004; Surdej et al. 2010; Yaitskova et al. 2005). The output is a measure of the second derivative of the phase, and there is good sensitivity to segment piston errors.

All four techniques (including Shack-Hartmann phasing) tested in the APE experiment were successful to some degree. Each technique has its advantages or disadvantages with respect

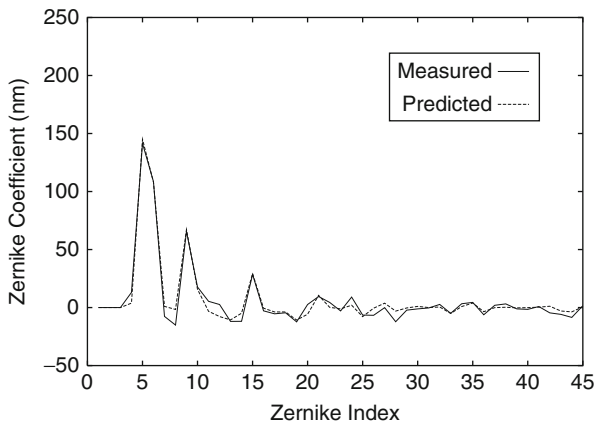
to pupil registration or determination of the pupil mapping, the fraction of the segment surface which is sampled, ability to determine segment tip/tilt errors simultaneously, sensitivity to rolled edges, and the ability to handle piston errors that are large compared to the wavelength of light. A full discussion of the relevant details may be found in the papers cited above.

10.3 Warping Harnesses

It is desirable to be able to adjust the segment figures (at least in the lower spatial frequencies) while the segments are in the telescope. In stressed mirror polishing, when the rounds are cut into hexagons, internal stresses are relieved, which tends to produce low spatial frequency aberrations that are very expensive to polish out at this late stage of the fabrication process. Also, errors in the in-plane positioning of segments cannot be corrected directly, as these degrees of freedom are not actively controlled; however, they can be corrected indirectly by adjusting the segment figures, since the predominant effect of these positioning errors is to cause apparent focus and astigmatism errors in the segments.

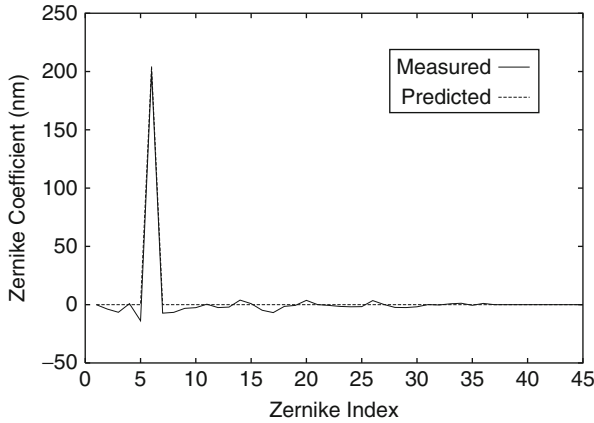
The in situ adjustment of segment figures is accomplished at Keck by means of warping harnesses, sets of leaf springs attached to the segment whiffletrees that can apply forces and moments to the backs of the segments so that they can be warped into the correct shapes. **◆** Figures 3-7 and **◆** 3-8 show the results of two warping harness experiments at Keck. In the first of these, the predicted changes that result from the application of a given force to a single warping beam were accurately confirmed. In the second, the astigmatism of one of the segments was successfully changed by 500 nm, to within the experimental errors of 15 nm.

Keck segments are normally measured and warped immediately after they are installed in the telescope, either initially or after aluminizing (and only at such times). The warping harnesses are typically not adjusted again until after the next aluminization cycle, typically 3 years later. Prior to warping, the Keck segments have a surface error of 98 nm (rms); after warping, this is reduced to 37 nm. The predicted performance for the warping harnesses is for a



◆ Fig. 3-7

Theoretical and observed differential response of a Keck telescope segment to the adjustment of a single leaf spring in the warping harness



■ Fig. 3-8

Results of an experiment to increase the astigmatism (Zernike polynomial 6) of a Keck segment by 200 nm via warping harness adjustment

post-warp rms of 25 nm; this includes both the theoretical performance of the springs and the errors associated with making Shack-Hartmann measurements through atmospheric turbulence. The discrepancy between theory and performance (25 vs. 37 nm) is apparently due to aliasing in the measurements and to the presence of significant non-Zernike aberrations, which the mathematical model for the segment surface error does not readily accommodate. Note that warping a known low spatial frequency aberration into a segment is not the same as (and is in general not as difficult as) flattening a segment with high spatial frequency aberrations present.

The Keck warping harness procedures involve manual adjustment of the warping harnesses. At the Gran Telescopio Canarias, this procedure can be done automatically, and this will also be the case for the giant segmented-mirror telescopes of the future. In principle, such warping harnesses can be exercised as often as several times a night to correct, for example, for temperature variations or gravity. This would require either continuous wavefront sensing or lookup tables to determine the necessary corrections.

10.4 Alignment of the Secondary Mirror

As is the case with any two mirror telescope, the secondary mirror of a segmented-mirror telescope must be aligned with respect to the primary in piston, tip, and tilt. (Precise centration of the secondary is typically not required, as the effect of decentering on the image is nearly degenerate with that of tip and tilt.) The complication for a segmented-mirror telescope is that the global rigid body parameters of the primary are not trivial to determine; at best, they can be represented as suitable averages over the corresponding segment quantities.

The basic idea underlying the alignment of the secondary in a segmented-mirror telescope can be understood as follows. In an otherwise perfect telescope, the defocus of the image due to a piston error δz (in microns) of the secondary is

$$AD = \delta z s (m^2 + 1)/F \quad (3.44)$$

where AD is the full diameter of image in arcsec, s is the image scale in arcsec per micron, m is the magnification of the secondary, and F is the overall focal ratio of the system. The corresponding global focus error can be expanded locally about the center of each segment. It follows immediately that each segment will have the same focus error to first order. Therefore the piston of the secondary can be adjusted until the average segment focus error is zero. The accuracy of the procedure is limited by atmospheric turbulence. At Keck, the average secondary mirror piston uncertainty is about $4\ \mu\text{m}$, as determined from the variation in a sequence of measurements.

The idea is similar for secondary tip/tilt. Schroeder (2000) gives the corresponding expression for the coma of the image due to a tilt of the secondary:

$$ATC = 3 \delta\alpha (m - 1) (1 + \beta) / (16F^2) \quad (3.45)$$

where $\delta\alpha$ is the secondary tilt angle, and β is the ratio of back focal distance to the focal length of the primary. Again, expanding the wavefront error about each segment center shows that each segment will have an apparent coma. In this case, each segment will also be astigmatic, with the astigmatism varying from segment to segment; the astigmatism in fact dominates. Thus, the secondary tip/tilt can be reconstructed from the ensemble of segment astigmatisms. The uncertainty in the secondary tip/tilt as determined at Keck is about 3 arcsec each in tip and tilt.

11 Other Segmented-Mirror Telescopes

In the last decade or so, several other large segmented-mirror ground-based telescopes have come on line, and the segmented James Webb Space Telescope is currently under construction. These telescopes are briefly described in the following sections.

11.1 Gran Telescopio Canarias

The Gran Telescopio Canarias (GTC) (Rodríguez Espinosa et al. 1999) is a 36-segment 10-m telescope with a design very similar to that of the Keck telescopes. The GTC, a project of Spain with the participation of Mexico and the University of Florida, is located on the island of La Palma in the Canary Islands and saw first light in 2007. The GTC incorporates several improvements over Keck, including the capability to continuously monitor the optical alignment and to adjust the segment figures remotely.

11.2 Hobby-Eberly Telescope and Southern African Large Telescope

The Hobby-Eberly Telescope (HET) at the McDonald Observatory in Texas (Barnes et al. 2000) is a segmented-mirror telescope designed to do spectroscopy and built for a fraction of the cost (about 20%) of a comparably sized segmented-mirror telescope like Keck or GTC. HET is joint project of the Pennsylvania State University and the University of Texas at Austin. Partner

institutions include Stanford University and two German institutions: Ludwig-Maximilians-Universität and Georg-August-Universität. Science operations began in October of 1999.

In order to achieve its dramatic cost reduction, the HET design incorporates a number of design innovations, the most dramatic of which is that the telescope is fixed in elevation and rotates only in azimuth. This means that the primary mirror is fixed with respect to gravity, a fact which greatly simplifies the support structure for the mirror. As the telescope tracks a star in elevation, its image moves with respect to the telescope across the focal surface. A star tracker, which contains a spherical aberration corrector and the science instrument, or the fore optics for fiber-fed instruments that reside below the telescope, follows the image. A large fraction of the sky is covered in this way. Observing times are limited compared to a conventional telescope that can move in altitude as well as azimuth; however, observations as long as 2.5 h are possible. Although the full HET array is 11.1 by 9.8 m, the effective diameter seen by the instruments is 9.2 m.

The HET primary mirror consists of five rings of hexagonal segments, each of which has a circumscribed diameter of 1.15 m. The central segment is included because it is not blocked by a secondary (the telescope is used only at prime focus). Thus, there are 91 segments and 273 actuators (three per segment). Unlike Keck, the HET segments are all spherical and identical, which means that the spacing between adjacent segments varies slightly from the center to the edge of the array. The HET segments are aligned not using starlight but rather using an artificial source and a Shack-Hartmann wavefront sensor located in a tower at the primary mirror center of curvature. The HET was originally designed without edge sensors, so that there was no active control of the primary mirror actuators. The telescope was realigned many times a night by rotating to the center of curvature alignment sensor. However, in the interest of observing efficiency, the telescope was retrofitted with an active control system, similar to the one used at Keck but with inductive, not capacitive, sensors. There are two sensors per intersegment edge for a total of 480, and the system updates once a minute. With the active control system in place, the HET segments only need to be aligned in tip/tilt at the beginning of each night. Because the telescope is not used for imaging, the segment piston tolerances are relatively loose and can be achieved by careful segment installation; the phasing procedures used at Keck are not necessary. A major upgrade to the telescope is planned in order to support an extensive dark-energy study known as HETDEX (Booth et al. 2006).

The Southern African Large Telescope (Buckley et al. 2004) is a southern hemisphere twin of the Hobby-Eberly, built near Sutherland, South Africa, by the South African Astronomical Observatory and a host of international partners. The baseline design was the same as that of the HET, but modifications were introduced to increase the field of view and to incorporate edge sensors and hence active control from the beginning. The telescope was inaugurated in November of 2005.

11.3 Large Area Multi-Object Spectrographic Telescope

The Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) (Cui et al. 2010) is a Schmidt telescope, built by China and devoted, as the name implies, to spectroscopy. The telescope has two segmented optics: a spherical primary mirror, roughly 6 m in diameter and consisting of 37 hexagonal segments of circumscribed diameter 1.1 m, and a corrector mirror, 5.72×4.4 m, consisting of 24 segments of the same size.

11.4 James Webb Space Telescope

Although the design considerations for a space-based telescope are much different than for a ground-based one, the James Webb Space Telescope (Sabelhaus and Decker 2004; Gardner et al. 2006), the successor to the Hubble Space Telescope, deserves mention in this review. The 6.5-m primary mirror of JWST will consist of 18 hexagonal segments in two rings. The segments are about the same size as those of Keck but made of light-weighted beryllium, not Zerodur. The motivation for the segmentation of JWST comes not only from fabrication considerations but also from those of launch: it must be both light-weight enough and compact enough to be launched later in this decade by an Ariane 5 rocket. The telescope will be launched in a folded configuration so that it can fit into the 4.5-m-diameter rocket and will be deployed on orbit.

JWST differs from ground-based segmented-mirror telescopes in that the primary mirror will not have an active control system; there are actuators but no sensors. There are several reasons for this. After initial deployment, there is little for the actuators to do since thermal perturbations to the system should be minor and of course the gravitational perturbations are nonexistent. In general, one wants to move actuators in a space-based system as seldom as possible in order to minimize the chance of an actuator failure, even one of which could jeopardize the entire mission. Additionally, a system of segment edge sensors and its electronics could create heat leaks which would compromise the cryogenic environment of this infrared telescope.

As noted in [Sect. 10.2](#), the coarse phasing of the JWST segments will be accomplished with a technique known as dispersed Hartmann sensing (DHS) (Sivaramakrishnan et al. 2003). In DHS, an array of prisms is placed over the intersegment edges at the location of a small image of the primary. The prisms disperse the light so that the wavelengths corresponding to constructive or destructive interference can be identified; the relative piston error can then be reconstructed from this information. The prisms also displace the interference patterns from one another on the detector so that multiple edges can be measured in parallel. Fine phasing will be accomplished using a variation of the Gerchberg-Saxton algorithm (Gerchberg and Saxton 1972; Redding et al. 2000) in which a Fourier transform-based procedure is used to reconstruct the wavefront from an out-of-focus image. Such techniques have not been used for ground-based telescopes because they are severely limited by atmospheric turbulence, but this is clearly not a problem in space.

12 Giant Segmented-Mirror Telescopes

There are three extremely large telescope projects currently in the design stages: the Giant Magellan Telescope (GMT), the Thirty Meter Telescope (TMT), and the European Extremely Large Telescope (E-ELT). All three projects involve large international collaborations, all have budgets in the range \$1 billion to €1 billion, and all expect to see first light around the end of the current decade. Because the GMT has only a small number of segments and because these are circular and of the lightweight honeycomb design, a detailed discussion of this telescope is left to [Chap. 4](#) of the current volume; only a brief description is given below. The emphasis here

is on the highly segmented TMT and to a lesser extent on the E-ELT. Further information on the E-ELT may be found in [Chap. 5](#) of the current volume.

12.1 GMT

The collaboration for the Giant Magellan Telescope (GMT) includes the five partners in the original Magellan project, which built the twin 6.5-m Magellan telescopes: the Carnegie Institution for Science, the University of Arizona, Harvard University, the Massachusetts Institute of Technology, and the University of Michigan, plus several other US universities as well as international partners. The GMT primary mirror will consist of seven circular 8.4-m segments and have an overall diameter of 24.5 m (<http://www.gmto.org/science-conceptu.html>). The hollow honeycomb segments are similar to those of the Large Binocular Telescope but are somewhat thinner and more flexible in order to accommodate potentially larger aberrations in the six off-axis segments, which are significantly asymmetric. The overall optical design is aplanatic Gregorian, with an exceedingly fast primary mirror focal ratio of $f/0.7$ and a final focal ratio of $f/8.0$.

Although the 700-mm-thick primary mirror segments are far stiffer than an 8-m meniscus mirror, some measure of active figure control is nevertheless required; this is supplied by an array of 165 actuators per segment. A ventilating system is used to reduce thermal gradients, since the glass has a nonzero coefficient of thermal expansion. The concave secondary mirror (3.2 m in diameter) is also segmented, with its seven segments conjugated to those of the primary. There will in fact be two secondary mirrors with this geometry: for diffraction-limited aberrations, an adaptive mirror whose segments are deformable thin face sheets, and, for seeing-limited observations, a fast steering mirror with rigid but independently moveable segments.

Wavefront information is supplied by a Shack-Hartmann wavefront sensor. In contrast to the electromechanical edge sensors used on Keck and GTC (and planned for TMT), the active alignment of the GMT segments in tip/tilt will be monitored by two (widely separated) optical sensors on each intersegment edge. According to current plans, the relative piston errors between segments will be measured using a modified phase-diversity scheme (Lloyd-Hart et al. 2006). The fact that the secondary mirror is conjugated to the primary means that the task of phasing the telescope segments can be divided between primary and secondary; for example, piston errors between segments could be sensed at the primary but corrected at the (more agile) secondary.

The GMT enclosure will be a cylinder approximately 65 m high. The telescope will rotate independently of the enclosure and will be able to observe at zenith angles from 0° to 65° . The GMT will be built at the existing Las Campanas Observatory in Chile, at an altitude of about 2,500 m. A site testing campaign will determine the optimal peak among several possible candidate sites at LCO.

12.2 TMT

The Thirty Meter Telescope (<http://www.tmt.org/sites/default/files/TMT-Construction-Proposal-Public.pdf>) represents a merger of three formerly independent segmented-mirror telescope projects: the University of California/Caltech California Extremely Large Telescope

(Nelson 2000), the National Observatory's Giant Segmented-Mirror Telescope (Strom et al. 2002), and Canada's Very Large Optical Telescope (Roberts et al. 2002). In addition to the University of California and Caltech (the original partners in the development of the Keck Observatory), the collaboration includes the Association of Canadian Universities for Research in Astronomy, and other international partners are expected to join.

For telescopes with monolithic mirrors, the financial rule of thumb is that overall cost scales with primary mirror diameter D as $D^{2.7}$. A TMT study concluded, by contrast, that for a segmented thirty meter telescope, the costs relative to the Keck capital investment should scale far more slowly. This, coupled with the D^4 dependence of the figure of merit for background limited observations, makes TMT an extremely good scientific value.

To be precise, 30.0 m is the diameter of the circumscribing circle around the primary mirror, which serves as the entrance pupil of the telescope. The optical design of TMT is Ritchey-Chretien (concave hyperbolic primary mirror and convex hyperbolic secondary), although, unlike most R-C telescopes, there is no Cassegrain focus. This is because an articulated tertiary mirror (2.45×3.51 m in diameter) not only folds the telescope beam but also rotates it to direct the light to the science instruments arrayed around the two Nasmyth platforms. This eliminates the need to reposition the very large instruments and also eliminates the need for a conventional (unfolded) Cassegrain focus. Like GMT, TMT will support both diffraction-limited and seeing-limited observations, but unlike GMT, there is only a single secondary mirror, and it is neither active nor adaptive. The primary mirror focal ratio is $f/1$, converted to $f/15$ by the 3.02 m diameter secondary mirror.

For adaptive optics observations, TMT will have a facility on the Nasmyth platform capable of feeding a 1 arcmin AO-compensated beam to multiple science instruments. This Narrow Field IR AO System (NFIRAOS) will utilize two deformable mirrors, conjugated to altitudes of 0 and 12 km.

The TMT primary mirror will consist of 492 hexagonal segments, each 0.715 m on a side (somewhat smaller than the 0.9 m Keck segments). There are $492/6 = 82$ different types of segments; one spare segment of each type will be provided. The Keck segments are identical regular hexagons when seen in projection. This means that the physical size of the outermost Keck segments is significantly larger than that of the innermost segments. The TMT design, by contrast, utilizes a scaling approach that keeps the physical size of all segments equal to within a few millimeters. As a consequence, the TMT segments are not seen as regular hexagons in projection but appear increasingly foreshortened with increasing radial distance from the center.

As was the case for Keck, the primary mirror control system for TMT will consist of three actuators per segment (a total of 1,476) and two edge sensors per intersegment edge (a total of 2,772). The TMT actuators are considerably more complicated than the Keck actuators in order to provide substantial vibration attenuation and damping in addition to achieving the desired positioning accuracy. Significant geometrical differences between the TMT and Keck edge sensors, described in [Sect. 9.3](#) above, help to reduce the sensor costs and to simplify segment exchanges. The large number of sensors makes it likely that at any given time, several of them will not be functional. At Keck, bad sensors are identified "by hand" and removed from the system. Computer algorithms that will do the identification automatically are being developed for TMT (Chanan and Nelson 2009).

In order to reach the 30-m diffraction limit, the TMT wavefront errors must be a factor of three smaller than the corresponding Keck errors. This presents a significant challenge for both segment fabrication and for optical alignment. The optical alignment system of TMT is still being designed, but it will be based on a Shack-Hartmann approach, similar to that used at Keck.

The TMT enclosure is not of the conventional dome/shutter design but rather involves a calotte configuration, consisting of a base, cap, and shutter. The base and cap form a spherical shell cut by a plane inclined at an angle of $\theta_{\text{cap}} = 32.5^\circ$, with respect to the horizontal. The base rotates about a vertical axis in the azimuth direction, and the cap rotates about an axis perpendicular to the plane that divides cap and base. The cap incorporates a circular aperture that can be positioned for observing at zenith angles from 0° to $2\theta_{\text{cap}} = 65^\circ$. The shutter structure, which rotates about the same axis, but independently of the cap, includes a plug to seal the aperture. Three rows of vents near the bottom of the base structure provide an open area of up to $1,700 \text{ m}^2$ for natural ventilation of the enclosure during observations.

The TMT site selection process was particularly thorough. On-site measurements with a variety of instruments, including differential image motion monitors and multi-aperture scintillation sensors, were undertaken at five prospective sites: Cerro Tolar, Cerro Armazones, and Cerro Tolonchar in Chile; San Pedro Martir in Mexico (Baja California); and Mauna Kea in Hawaii, for a period of 1–2 years at each site. The atmospheric data were supplemented by existing long-term data sets and also by computational fluid dynamics simulations. Other issues considered in the selection included construction and operating costs; cultural, environmental, and land use issues; labor force issues; proximity to astronomers and to astronomical infrastructure; geological conditions; and other factors. Based on all of these considerations, in July of 2009, the TMT Board selected Mauna Kea as the preferred site for TMT.

12.3 E-ELT

The European Extremely Large Telescope (E-ELT) (http://www.eso.org/sci/facilities/eelt/docs/e-elt_constrproposal.pdf) is a project of the 14-member-nation European Southern Observatory. The E-ELT diameter of 42 m will give it twice the area of the Thirty Meter Telescope.

The optical design of the E-ELT is an anastigmat with three powered mirrors and two fold-flats. The $f/1$ primary mirror (M1) consists of 984 segments with a nominal circumscribed diameter of 1.45 m, very similar to those of TMT. The convex secondary (M2) is 6 m in diameter. A 4.2-m concave and mildly aspheric tertiary (M3) relays light to a flat 2.5-m deformable mirror (M4) with some 5,000 degrees of freedom and then a 2.7-m field stabilizing mirror (M5). The primary mirror focal ratio is $f/1$ and the final Nasmyth focal ratio is $f/17.7$. Besides the two Nasmyth foci, there is a gravity-invariant focus that is fed by an additional fold flat (M6) and an $f/60$ Coude focus.

The enclosure for the E-ELT will be a hemispherical dome with a design that is fairly conventional except for its 100 m diameter; the design for the shutter has not yet been finalized.

In April 2010, the ESO Council selected Cerro Armazones (altitude 3,060 m) as the baseline site for the E-ELT. Cerro Armazones is in the central part of Chile's Atacama Desert, about 20 km from the VLT on Cerro Paranal.

Acknowledgments

The preparation of this chapter was supported in part by the Thirty Meter Telescope Corporation. The authors gratefully acknowledge the support of the TMT partner institutions: the Association of Canadian Universities for Research in Astronomy (ACURA), the California

Institute of Technology, and the University of California. This work was supported as well by the Gordon and Betty Moore Foundation, the Canada Foundation for Innovation, the Ontario Ministry of Research and Innovation, and the National Research Council of Canada.

References

- Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A., & Sorensen, D. 1999, *LAPACK User's Guide* (3rd ed.; Philadelphia: Society for Pure and Applied Mathematics)
- Andersen, T., Ardeberg, A., Beckers, J., Gontcharov, A., Owner-Petersen, M., & Riewaldt, H. 2004, Euro 50, in *Proceedings of the Second Backaskog Workshop on Extremely Large Telescopes*, eds. A. Ardeberg and T. Andersen, (Bellingham: SPIE). Proc SPIE 5382, 169–181
- Barnes, T. G., Adams, M. T., Booth, J. A., Cornell, M. E., Gaffney, N. I., Fowler, J. R., Hill, G. J., Hill, G. M., Nance, C. E., Piche, F., Ramsey, L. W., Ricketts, R. L., Spiesman, W. J., & Worthington, P. T. 2000, Commissioning experience with the 9.2-m Hobby-Eberly telescope, in *Telescope Structures, Enclosures, Controls, Assembly/Integration/Validation, and Commissioning*, ed. T. A. Sebring & T. Andersen (Bellingham: SPIE). Proc. SPIE, 4004, 14–25
- Beckers, J. M., Ulich, B. L., & Williams, J. T. 1982, MMT – the first of the advanced technology telescopes. Proc. SPIE, 332, 2–8
- Booth, J. A., MacQueen, P. J., Good, J. M., Wesley, G. L., Hill, G. J., Palunas, P., Segura, P. R., & Calder, R. E. 2006, The wide field upgrade for the Hobby-Eberly telescope. Proc. SPIE, 6267, 62673W-1–62673W-11
- Braunecker, B., Hentschel, R., & Tiziani, H. J. 2008, *Advanced Optics Using Aspherical Elements* (Bellingham: SPIE), SPIE Press Monograph Vol. PM 173
- Buckley, D. A. H., Meiring, J. G., Swiegers, J., & Swart, G. P. 2004, Many segments and few dollars: SALT solutions for ELTs? Proc. SPIE, 5382, 245–256
- Burgarella, D., Dohlen, K., Ferrari, M., Zamkotsian, F., Hammer, F., Sayede, F., & Rigaut, R. 2002, Large petal telescope for the next-generation Canada-France-Hawaii Telescope, in *Future Giant Telescopes*, ed. J. R. P. Angel & R. Gilmozzi (Bellingham: SPIE). Proc. SPIE, 4840, 93–103
- Chanan, G. A. 1988, Design of the Keck Observatory alignment camera, in *Precision Instrument Design*, ed. T. C. Bristow & A. E. Hathaway (Bellingham: SPIE). Proc. SPIE, 1036, 59–70
- Chanan, G., & Nelson, J. 2009, Algorithm for the identification of malfunctioning sensors in the control systems of segmented mirror telescopes. Appl. Opt., 48, 6281–6289
- Chanan, G., & Troy, M. 1999, Strehl ratio and modulation transfer function for segmented mirror telescopes as functions of segment phase error. Appl. Opt., 38, 6642–6647
- Chanan, G. A., Troy, M., Dekens, F. G., Michaels, S., Nelson, J., Mast, T., & Kirkman, D. 1998, Phasing the mirror segments of the Keck telescopes: the broadband phasing algorithm. Appl. Opt., 37, 140–155
- Chanan, G., Troy, M., & Sirko, E. 1999, Phase discontinuity sensing: a method for phasing segmented mirrors in the infrared. Appl. Opt., 38, 704–713
- Chanan, G. A., Ohara, C., & Troy, M. 2000, Phasing the mirror segments of the Keck telescopes: the narrowband phasing algorithm. Appl. Opt., 39, 4706–4714
- Chanan, G., MacMartin, D., Nelson, J., & Mast, T. 2004, Control and alignment of segmented-mirror telescopes: matrices, modes, and error propagation. Appl. Opt., 43, 1223–1232
- Chevillard, J.-P., Connes, P., Cuisenier, M., Friteau, J., & Marlot, C. 1977, Near infrared astronomical light collector. Appl. Opt., 16, 1817–1833
- Chueca, S., Reyes, M., Schumacher, A., & Montoya, L. 2008, DIPSI: measure of the tip-tilt with a diffraction phase sensing instrument. Proc. SPIE, 7012, 701213-1–701213-11
- Cohen, R., Mast, T., & Nelson, J. 1994, Performance of the W. M. Keck telescope active mirror control system, in *Advanced Technology Optical Telescopes*, ed. V. L. M. Stepp (Bellingham: SPIE). Proc. SPIE, 2199, 105–116
- Cui, X., Su, D.-Q., Wang, Y.-N., Li, G., Lui, G., Zhang, Y., & Li, Y. 2010, The optical performance of LAMOST telescope. Proc. SPIE, 7733, 773309-1–773309-8
- Dierickx, P., Beckers, J. L., Brunetto, E., Conan, R., Fedrigo, E., Gilmozzi, R., Hubin, N. N., Koch, F., Lelouarn, M., Marchetti, E., Monnet, G. J., Noethe, L., Quattri, M., Sarazin, M. S., Spyromilio, J., & Yaitskova, N. 2002, The eye of the

- beholder: designing the OWL, in *Future Giant Telescopes*, ed. J. R. P. Angel & R. Gilmozzi (Bellingham: SPIE). *Proc. SPIE*, 4840, 151–170
- Esposito, S., Pinna, E., Puglisi, A., Tozzi, A., & Stefanini, P. 2005, Pyramid sensor for segmented mirror alignment. *Opt. Lett.*, 30, 2572–2574
- Fried, D. L. 1966, Optical resolution through a randomly inhomogeneous medium for very long and very short exposures. *JOSA*, 56, 1372–1379
- Gardner, J. P., et al. 2006, The James Webb Space Telescope. *Space Sci. Rev.*, 123, 485–606
- Gerchberg, R. W., & Saxton, W. O. 1972, A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35, 237–246
- Gilmozzi, R., & Spyromilio, J. 2008, The 42m European ELT: status, in *Ground-Based and Airborne Telescopes II*, ed. L. Stepp & R. Gilmozzi (Bellingham: SPIE), 5662, 701219-1–701219-10
- Golub, G., & van Loan, C. 1996, *Matrix Computations* (3rd ed.); London: The Johns Hopkins University Press)
- Gonte, F. Y., Yaitskova, N., Dierickx, P., Karban, R., Courteville, A., Schumacher, A., Devaney, N., Esposito, S., Dohlen, K., Ferrari, M., & Montoya, L. 2004, APE: a breadboard to evaluate new phasing technologies for a future European giant optical telescope. *Proc. SPIE*, 5489, 1184–1191
- Horn D'Arturo G. 1955, *Pubblicazioni dell'Osservatorio Astronomico Universitario di Bologna*, VI, 6
http://www.eso.org/sci/facilities/eelt/docs/e-elt_constrproposal.pdf
<http://www.gmto.org/science-conceptu.html>
<http://www.tmt.org/sites/default/files/TMT-Construction-Proposal-Public.pdf>
- Jared, R. C., Arthur, A. A., Andreae, S., Biocca, A., Cohen, R. W., Fuentes, J. M., Franck, J., Gabor, G., Llacer, J., Mast, T., Meng, J., Merrick, T., Minor, R., Nelson, J., Orayani, M., Salz, P., Schaefer, B., & Witebsky, C. 1990, The W. M. Keck telescope segmented primary mirror active control system. *Proc. SPIE*, 1236, 996–1008
- Johns, M. 2008, The Giant Magellan Telescope (GMT). *Proc. SPIE*, 6986, 698603-1–698603-12
- Lloyd-Hart, M., Angel, R., Milton, N. M., Rademacher, M., & Codona, J. 2006, Design of the adaptive optics system for the GMT. *Proc. SPIE*, 6272, 62720E-1–62720E-12
- Lofdahl, M., & Eriksson, H. 2001, An algorithm for resolving 2π ambiguities in interferometric measurements by use of multiple wavelengths. *Opt. Eng.*, 40, 984–990
- Lubliner, J., & Nelson, J. 1980, Stressed mirror polishing: a technique for producing non-axisymmetric mirrors. *Appl. Opt.*, 19, 2332–2340
- MacMartin, D. G., & Chanan, G. A. 2002, Control of the California extremely large telescope primary mirror, in *Future Giant Telescopes*, ed. J. R. P. Angel & R. Gilmozzi (Bellingham: SPIE). *Proc. SPIE*, 4840, 69–80
- MacMartin, D. G., & Chanan, G. 2004, Measurement accuracy in control of segmented-mirror telescopes. *Appl. Opt.*, 43, 608–615
- Mast, T., & Nelson, J. 1982, Figure control for a fully segmented primary mirror. *Appl. Opt.*, 21, 2631–2641
- Mast, T., & Nelson, J. 2000, Segmented mirror control system hardware for CELT, in *Optical Design, Materials, Fabrication, and Maintenance*, ed. P. Dierickx (Bellingham: SPIE). *Proc. SPIE*, 4003, 226–240
- Meng, J. D., Minor, R., Merrick, T., & Gabor, G. 1990, Position control of the mirror figure control actuator for the Keck Observatory ten meter primary mirror. *Proc. SPIE*, 1236, 1018–1022
- Minor, R. H., Arthur, A. A., Gabor, G., Jackson, H. G., Jr., Jared, R. C., Mast, T. S., & Schaefer, B. A. 1990, Displacement sensors for the primary mirror of the W. M. Keck telescope. *Proc. SPIE*, 1236, 1009–1017
- Montoya, L. 2004, *Applications de l'interferometrie de Mach-Zehnder au cophasage des grands telescopes segmentes*. Ph.D. Thesis, Universite de Provence
- Nelson, J. E. 2000, Design concepts for the California Extremely Large Telescope (CELT). *Proc. SPIE*, 4004, 282–289
- Nelson, J., & Sanders, G. H. 2006, TMT status report, in *Ground-Based and Airborne Telescopes*, ed. L. M. Stepp (Bellingham: SPIE). *Proc. SPIE*, 6267, 745–761
- Nelson, J., & Temple-Raston, M. 1982, The off-axis expansion of conic surfaces. University of California TMT Report No. 91
- Nelson, J., Gabor, G., Hunt, L., Lubliner, J., & Mast, T. 1980, Stressed mirror polishing: fabrication of an off-axis section of a paraboloid. *Appl. Opt.*, 19, 2341–2352
- Nelson, J., Lubliner, J., & Mast, T. 1982, Telescope mirror supports: plate deflections on point supports. *Proc. SPIE*, 332, 212–228
- Nelson, J. E., Mast, T. S., & Faber, S. M. 1985, The design of the Keck observatory and telescope. Keck Observatory Report No. 90 (Berkeley: Keck Observatory Science Office)
- Noethe, L. 2002, Active optics in modern large optical telescopes. *Prog. Opt.*, 43, 1–69

- Noll, R. J. 1976, Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.*, 66, 207–211
- Papadogiannis, A. S., Papadogianni, N. S., Carabelas, A., Tsitomeneas, S., Kyraggelos, P., & Chondros, T. G. 2009, The mirror weapon in Archimedes' era. (Dordrecht: Springer) *Proc. EUCOMES08*, 29–36
- Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. 1989, *Numerical Recipes: The Art of Scientific Computing* (1st ed.; New York: Cambridge University Press)
- Redding, D. C., Basinger, S. A., Cohen, D., Lowman, A. E., Shi, F., Bely, P. Y., Bowers, C. W., Burg, R., Burns, L. A., Davila, P. S., Dean, B. H., Mosier, G. E., Norton, T. A., Petrone, P., Perkins, B. D., & Wilson, M. 2000, Wavefront control for a segmented deployable space telescope. *Proc. SPIE*, 4013, 546–558
- Roberts, S. C., Morbey, C. L., Crabtree, D. R., Carlberg, R., Crampton, D., Davidge, T. J., Fitzsimmons, J. T., Gedig, M. H., Halliday, D. J., Hesse, J. E., Herriot, R. G., Oke, J. B., Pazder, J. S., Szeto, K., & Veran, J.-P. 2002, Canadian very large optical telescope technology studies, in *Future Giant Telescopes*, ed. J. R. P. Angel & R. Gilmozzi (Bellingham: SPIE). *Proc. SPIE*, 4840, 104–115
- Roddier, F. 1988, Curvature sensing and compensation: a new concept in adaptive optics. *Appl. Opt.*, 27, 1223–1225
- Rodriguez Espinosa, J. M., Alvarez, P., & Sanchez, F. 1999, The GTC: an advanced 10m telescope for the ORM. *Astrophys. Space Sci.*, 263, 355–360
- Sabelhaus, P., & Decker, J. 2004, An overview of the James Webb Space Telescope (JWST) project. *Proc. SPIE*, 5487, 550–563
- Schroeder, D. J. 2000, *Astronomical Optics* (San Diego, CA: Academic)
- Schumacher, A., Devaney, N., & Montoya, L. 2002, Phasing segmented mirrors: a modification of the Keck narrow-band technique and its application to extremely large telescopes. *Appl. Opt.*, 41, 1297–1307
- Shi, F., Chanan, G., Ohara, C., Troy, M., & Redding, D. C. 2004, Experimental verification of dispersed fringe sensing as a segment phasing technique using the Keck telescope. *Appl. Opt.*, 43, 4474–4481
- Sivaramakrishnan, A., Makidon, R. B., Acton, D. S., & Shi, F. 2003, Coarse phasing JWST using dispersed fringe sensing and dispersed Hartmann sensing during commissioning. *Space Telescope Science Institute Technical Memorandum STSCI-JWST-TM-2003-0022A*, Baltimore
- Strom, S. E., Stepp, L. M., & Gregory, B. 2002, Giant segmented mirror telescope: a point design based on science drivers, in *Future Giant Telescopes*, ed. J. R. P. Angel & R. Gilmozzi (Bellingham: SPIE). *Proc. SPIE*, 4840, 116–128
- Surdej, I., Yaitskova, N., & Gontje, F. 2010, On-sky performance of the Zernike phase contrast sensor for the phasing of segmented telescopes. *Appl. Opt.*, 49, 4053–4063
- Troy, M., & Chanan, G. 2003, Diffraction effects from giant segmented-mirror telescopes. *Appl. Opt.*, 42, 3745–3753
- West, S. C., Callahan, S., Chaffee, F. H., Davidson, W. B., Derigne, S. T., Fabricant, D. G., Foltz, C. B., Hill, J. M., Nagel, R. H., & Poyner, A. D. 1997, Toward first light for the 6.5-m MMT telescope. *Proc. SPIE*, 2871, 38–48
- Yaitskova, N., Dohlen, K., Dierickx, P., & Montoya, L. 2005, Mach-Zehnder interferometer for piston and tip-tilt sensing in segmented telescopes: theory and analytical treatment. *JOSA A*, 22, 1093–1105
- Ziad, A., Schoeck, M., Chanan, G. A., Troy, M., Dekany, R., Lane, B. F., Borgnino, J., & Martin, F. 2004, Comparison of measurements of the outer scale of turbulence by three different techniques. *Appl. Opt.*, 43, 2316–2324

4 Honeycomb Mirrors for Large Telescopes

*John Hill*¹ · *Hubert Martin*² · *Roger Angel*²

¹Large Binocular Telescope Observatory, University of Arizona, Steward Observatory, Tucson, AZ, USA

²Department of Astronomy, University of Arizona, Steward Observatory, Tucson, AZ, USA

1	<i>Introduction</i>	139
1.1	Large Telescopes Using Honeycomb Mirrors	139
1.2	Experience of the MMT	140
1.3	Early Casting Experiments	140
1.4	Development of SOML and Hextek	141
1.5	Requirements for Ground-Based Telescope Primary Mirrors	141
1.5.1	Large Collecting Area	142
1.5.2	Image Quality	142
1.5.3	Diffraction-Limited Performance in the Infrared	142
1.5.4	Lightweight Optics	143
1.5.5	Structural Stiffness of the Optics	143
1.5.6	Low Thermal Inertia (Mirror Seeing)	144
1.5.7	Low CTE and Fast Thermal Equilibrium	144
1.5.8	Affordable	145
1.5.9	Compact Focal Lengths	145
2	<i>Casting Borosilicate Honeycomb Mirror Blanks</i>	145
2.1	Key Concepts	145
2.2	Process Description	146
2.2.1	Mold Assembly	146
2.2.2	Furnace Description	148
2.2.3	Casting	149
2.2.4	Annealing and Cooling	149
2.2.5	Mold Removal and Handling	151
2.3	Why E6 Borosilicate Glass?	152
2.4	Blank Dimensions and Properties	153
2.4.1	Honeycomb Sandwich Design	153
2.4.2	Dimensions	153
3	<i>Optical Fabrication and Testing of Large, Aspheric Mirrors</i>	153
3.1	Overview of Fabrication and Testing	154
3.2	Accuracy Requirements	155

3.3	Polishing of Large Mirrors	159
3.4	Measurement of Large Mirrors	161
4	<i>Experience in Operating Astronomical Telescopes</i>	166
4.1	SDSS 2.5 Meter	166
4.2	Lennon 1.8 Meter	166
4.3	APO 3.5 Meter	166
4.4	WIYN 3.5 Meter	166
4.5	SOR 3.5 Meter	167
4.6	MMT 6.5 Meter	167
4.7	Magellan 6.5 Meter	168
4.8	LBT 2 × 8.4 Meter	168
4.8.1	Support and Positioning of Primary Mirrors	170
4.8.2	Active Optics and Wavefront Sensing	171
4.8.3	Ventilation and Thermal Management	173
4.8.4	Reflective Coating	173
4.8.5	Handling and Transportation	174
4.8.6	Adaptive Optics	174
4.8.7	Phased Array Imaging	176
4.8.8	Partners	178
5	<i>Future Telescopes Under Development</i>	178
5.1	LSST	178
5.2	GMT	178
5.3	SASIR	179
	<i>Acknowledgements</i>	179
	<i>References</i>	180

Abstract: This chapter deals with the design, fabrication, and use of the borosilicate glass honeycomb mirrors which are being produced at the University of Arizona's Steward Observatory Mirror Laboratory. These mirrors are a core technology for the whole telescope, and a number of telescopes are now operational using these primary mirrors. The mirrors contribute to the telescope design because of their light weight, their high stiffness, and their short thermal time constant. The light weight of the primary mirrors helps to keep the weight of the entire telescope low and to maximize the structural performance. The ability to circulate air through the glass honeycomb structure allows control of local seeing in the telescope environment. The honeycomb sandwich is formed by spin casting borosilicate glass into a ceramic fiber mold. The Mirror Lab has previously produced three 3.5-m mirrors, three 6.5-m mirrors, and two 8.4-m mirrors which are now operating successfully in telescopes. Results are highlighted from these telescopes with emphasis on the Large Binocular Telescope with two 8.4 m primaries. Excellent results have been obtained with adaptive secondary mirrors in combination with the honeycomb primary mirrors. Two additional 6.5-m mirrors and two additional 8.4-m mirrors have also been cast and are in various stages of production for other projects including the first off-axis segment for the future Giant Magellan Telescope. An additional key technology for large telescopes is the ability to fabricate high-precision primary optics with short focal lengths in order to keep the telescope structure and enclosure compact. The stressed lap allows efficient polishing of these fast conic surfaces by actively adjusting its shape as it strokes across the mirror.

Keywords: Active optics, Adaptive optics, Beam combination, Borosilicate glass, Diffraction limit, Future telescopes, Gas fusion, Honeycomb sandwich, Lightweight optics, Mirror seeing, Mirror support, Off-axis paraboloid, Optical testing, Phased Array imaging, Polishing, Primary mirror, Prime focus, Spin casting, Strehl ratio, Stressed lap, Telescope

List of Abbreviations: *MMT*, Multiple Mirror Telescope; *LBT*, Large Binocular Telescope; *CTE*, Coefficient of thermal expansion; *SOML*, Steward Observatory Mirror Laboratory (Mirror Lab); *SDSS*, Sloan Digital Sky Survey; *VATT*, Vatican Advanced Technology Telescope (Alice P. Lennon Telescope); *APO*, Apache Point Observatory; *WIYN*, Wisconsin Indiana Yale NOAO; *GMT*, Giant Magellan Telescope; *LSST*, Large Synoptic Survey Telescope; *NOAO*, National Optical Astronomy Observatory; *IMACS*, Inamori Magellan Areal Camera; *HST*, Hubble Space Telescope; *PSF*, Point spread function; *RTV*, Room temperature vulcanizing silicone; *IRAF*, Interactive Reduction and Analysis Facility; "*km, m, cm, mm, μ m, nm*", Metric distances; *FWHM*, Full width at half maximum; *P*, Poise (unit of viscosity); *SASIR*, Synoptic All-Sky Infrared Imaging Survey; *CCD*, Charge coupled device; *DSP*, Digital signal processor

1 Introduction

1.1 Large Telescopes Using Honeycomb Mirrors

Honeycomb mirrors were developed at the University of Arizona to enable the construction of large telescopes with excellent performance at moderate cost. The performance is a result of the mirrors' high stiffness against gravity and wind, their rapid response to changing ambient temperature, and the fact that monolithic mirrors automatically provide a smooth wavefront. The honeycomb mirrors help control telescope costs because of their low mass, low sensitivity to support forces, and the fact that short focal lengths were adopted for all telescopes using

honeycomb mirrors. The operating large telescopes using honeycomb mirrors are the 6.5-m MMT (↗ Sect. 4.5), the two 6.5-m Magellan telescopes (↗ Sect. 4.7), and the Large Binocular Telescope (↗ Sect. 4.8) with two 8.4-m mirrors. These and other telescopes are described below in Experience in Operating Astronomical Telescopes. Additional telescopes described in Future Telescopes Under Development will also use honeycomb mirrors. The Large Synoptic Survey Telescope (↗ Sect. 5.1) will use an 8.4-m honeycomb substrate for its combined primary and tertiary mirrors. The Giant Magellan Telescope (↗ Sect. 5.2) will use seven 8.4-m honeycomb segments to form its 25-m primary mirror.

1.2 Experience of the MMT

The original Multiple Mirror Telescope (↗ Sect. 4.5) was built in the late 1970s using six 1.8-m diameter mirror blanks which were surplus satellite optics of honeycomb construction (Beckers and Williams 1980; Beckers et al. 1981). One of the favorable surprises from the early observations with the MMT on Mt. Hopkins in 1979 was that the telescope sometimes gave superb image quality (<0.5 arcsec) (Beckers et al. 1982) which was not expected from typical ground-based telescopes at that time (Woolf 1982a). It was quickly recognized by Woolf (1979) that the light weight honeycomb primary mirrors and the light weight telescope structure were significant contributors to the reduced amount of local dome seeing (image blurring due to air turbulence) which was observed. The concept of honeycomb mirrors in large telescope dates back to Ritchey (1928) who recognized the potential performance gains of light weight optics some 50 years earlier. The success of the original MMT led Angel and Woolf to start considering large telescope designs which took advantage of light weight honeycomb mirrors and to direct attention to seeing management in the telescope and enclosure designs (Angel and Woolf 1980; Woolf 1982b).

1.3 Early Casting Experiments

In the fall of 1980, Angel and Hill (1981) undertook a series of experiments in the fabrication of honeycomb primary mirrors from borosilicate glass. The idea was to achieve the excellent imaging performance seen in the MMT, while using a simpler and less expensive glass technology than the fused silica used in the MMT honeycomb mirrors. Borosilicate glass is relatively inexpensive, is easy to form into complex shapes at modest temperatures, and already had significant astronomer experience in the Hale 5-m telescope on Mt. Palomar with its borosilicate ribbed structure produced by Corning Glass in 1935 (McCauley 1934). The initial experiments attempted to replicate the fused eggcrate structure of the MMT primary mirrors in borosilicate glass. The honeycomb rib structure was easy to replicate, but fusing together the edges of flat plates to make facesheets larger than a single sheet of glass was problematic. Eventually, the experimental path led to turning up the furnace temperature to $1,200^{\circ}\text{C}$ and melting broken pieces of borosilicate glass into a ceramic fiber mold. The higher temperature allowed the glass pieces to melt together and to flow into the mold to create the continuous honeycomb sandwich structure including the facesheet. This casting technology opened the path to making a borosilicate honeycomb blank of nearly any diameter without constraint by the size of available glass sheets (Angel et al. 1981; Angel and Hill 1982; Hill and Angel 1983; Angel and Hill 1984).

1.4 Development of SOML and Hextek

Some early experiments with fusing close-packed glass tubes together between plates to form a honeycomb structure were commercialized by Hextek Inc. (Parks et al. 1990). Hextek has been producing borosilicate honeycomb mirrors with this gas fusion technology in the 1 m size range for more than 20 years. Slight air pressure is used to inflate glass tubes so that they morph into a honeycomb as they are fused between facesheets. The resulting flat blanks can be slumped to the desired radius of curvature. The experiments in casting borosilicate glass into molds at higher temperature led to the formation of the Steward Observatory Mirror Laboratory (SOML) located under the stands of the University of Arizona football stadium in Tucson. The initial casting experiments produced two 1.8-m f/1.75 honeycomb blanks in 1983 (Angel and Hill 1984) followed by the spin casting of the 1.8-m f/1.0 primary for the VATT telescope in 1985 (Goble et al. 1985). Construction of the present laboratory began in 1984 as part of a program to produce several 3.5-m honeycomb mirrors as a stepping stone to larger sizes (Angel et al. 1990). These 3.5-m mirrors were used in the APO, WIYN, and SOR telescopes discussed below in Experience in Operating Astronomical Telescopes. The Mirror Lab has specialized in the development and production (both casting and polishing) of large honeycomb mirrors from 3.5- to 8.4-m diameter.

1.5 Requirements for Ground-Based Telescope Primary Mirrors

A key step in the development of the borosilicate honeycomb mirrors was to understand what design parameters were required for mirrors in large ground-based telescopes (Angel and Woolf 1984). Astronomers always want a large collecting area and sharp images in order to have a crisp view of extremely faint distant objects (Pease 1926). Historically, each significant step in collecting area has also resulted in unexpected new astronomical discoveries along with the expected gains in light-gathering power. To constrain our design, it is necessary to peel back another level of engineering detail and see what those basic astronomer requirements entail for practical large telescope mirrors. At the same time, the technological solutions for large optics must be explored to see how they mesh with the astronomical requirements. Those are the iterative trade-offs needed for a successful telescope design. The successful design must meet the astronomical needs while being produced with practical technology. Note that there are other mirror technology solutions for large telescopes, developed around the same time that are described in adjacent chapters on [Chap. 3](#) and [Chap. 5](#).

The following parameters must be optimized for large ground-based telescopes: maximum light-collecting area, maximum light collection efficiency, and sharpest image quality. At the same time, the engineering solutions must fight the effects of varying gravity forces as the telescope moves around the sky and the effects of wind and other stray forces bending the mirrors. The technology must also deal with the effects of a changing temperature environment in order to minimize local mirror seeing and other disturbances so that the delivered image quality will be limited only by the turbulence of the free atmosphere. The importance of temperature management was not widely recognized in earlier generations of telescopes built prior to 1975 (Woolf 1979). The design of the optics must have minimum impact on the telescope structure, the telescope enclosure, and the resulting cost so that the entire observatory is practical to build.

Those considerations have led to the following specific items to be optimized during the design and production of the borosilicate honeycomb mirrors for ground-based telescopes.

1.5.1 Large Collecting Area

The signal-to-noise ratio for observing distant objects with a telescope almost always depends on the total number of photons collected. The need for photons is particularly critical for high-resolution spectroscopy where the total number of photons is dispersed among many spectral resolution elements, and observations are frequently limited by readout noise. Optical/infrared telescopes with the largest collecting areas are built on the ground rather than in space because the area (size of the telescope) is not limited by the launch capability. For some observations (those which are photon starved on bright objects), the ground-based telescopes are able to compete with the lower background in space-based telescopes by having larger collecting area.

1.5.2 Image Quality

All telescopes, especially the largest telescopes, must have excellent image quality. They cannot simply be “light buckets,” or smaller telescopes with better images will be able to outperform them by detecting faint astronomical objects against a background. Because of the large size of the primary mirrors and their important light collecting function in the telescope, the optical surface figure specification of the primary optics is critical to the cost, schedule and performance of the telescopes. The goal is always the best possible image quality, but there is risk in asking for too much. Overspecifying the figure for polishing the optics may drain resources from other parts of the telescope. The figure of the mirrors for ground-based telescopes is set to match the performance of the atmosphere on the best nights as opposed to being fully diffraction-limited in visible light (Hill 1990). Most large telescopes of the current generation adopt a similar strategy for the optical specification as used by the Keck 10-m telescopes (Keck Observatory Project Office 1985). This strategy puts the atmosphere as the limiting factor for image quality all the time. This also means that any adaptive optics correction which can compensate the atmospheric wavefront can also compensate residual static errors in the telescope optics. The result of this specification strategy is that the small to mid-scale (10–20 cm) surface errors are the most critical. The detailed requirements for polishing are discussed in more detail below in [Sect. 3.2](#).

1.5.3 Diffraction-Limited Performance in the Infrared

A key assumption in designing telescopes with large apertures (larger than 4 m for the purposes of this discussion) is that some observations will be made in the infrared where the diffraction-limit sets the ultimate imaging performance. The effects of atmospheric turbulence on seeing decrease as the wavelength increases, so diffraction takes over as the limit on image quality in the mid-infrared (Dierickx et al. 1990). At shorter wavelengths, technology limitations may restrict the image quality so that multiple smaller apertures may be the most effective solution for a particular telescope. Diffraction-limited imaging (with adaptive optics in the near infrared and without the need for adaptive optics at longer mid-infrared

wavelengths) forces the telescope to have sections of continuous phased aperture in the 6–10-m size range. Imaging limited by atmospheric turbulence at shorter visual wavelengths can be done more economically with arrays of smaller nonphased apertures (with the exception of speckle interferometry which requires the implicit resolution of larger apertures) (Kaiser et al. 2002). Developments in telescope design and adaptive optics have made this case even stronger with future telescopes pushing to diffraction-limited performance over 25–40-m apertures such as those described below in [Sect. 5](#). Some multiple mirror or multiple telescope designs are able to improve the single-aperture diffraction limit by combining the light from multiple apertures. The Large Binocular Telescope ([Sect. 4.8](#)) discussed below is an example of combining the light from two 8.4-m apertures to achieve additional resolution over a 22.65-m diffraction baseline. For some observations, the ground-based telescopes are able to compete with the lower background in space-based telescopes by having larger apertures and better diffraction-limited resolution. This advantage has only recently been realized on the ground with the advent of adaptive optics as discussed in [Sect. 4.8.6](#) to minimize the effects of atmospheric turbulence.

1.5.4 Lightweight Optics

The mirror substrate should be light weight – a restriction that becomes more demanding as the size and resulting mass of the telescope increases. Scaling up classical mirror substrates with aspect ratios (diameter/thickness) of 6:1 quickly leads to 8-m primary mirrors that are slabs of glass over a meter thick with weights approaching 100 tons. Many people have recognized that these classical designs with heavy slabs of glass are impractical even when executed in zero-expansion materials. There are serious problems with mirror seeing when such a thick slab of glass is always out of temperature equilibrium (thermal time constant of 1 week) with the mountaintop environment. The weight of the classical blank also drives the entire telescope structure to be far too heavy and expensive. Lightweight optics allow the optical support system and the entire telescope structure to be much less massive at a given level of optical performance.

1.5.5 Structural Stiffness of the Optics

Having optics that are light weight is not sufficient. The telescope mirrors must also be stiff enough to resist disturbing forces and maintain their precise shapes. Of course, the function of the mirror substrate is to support and stabilize the thin metal layer which is the actual reflecting surface. The honeycomb sandwich structure is an opportunity to preserve the inherent stiffness against gravity of the 6:1 aspect ratio of earlier mirrors, while reducing the total weight by a factor of order 5. The present borosilicate honeycomb blanks are designed to have a stiffness/weight ratio similar to a thick solid mirror. That gives them a stable figure against the forces of wind and gravity. As mirrors grow to even larger diameters (and become more difficult to shield from the wind), it is the wind forces that provide the ultimate limit to how much the stiffness can be reduced by light weighting.

All ground-based telescope optics must be carefully supported to minimize flexure under gravity loading. The common approach is to provide a distributed support force to float the optic against gravity. The honeycomb substrates have been optimized to reproduce the original figure with fairly simple mechanical supports. For more flexible mirrors, the tolerances on the support

forces become tighter. Thus, there is a practical trade-off between the stiffness and weight of the substrate, on the one hand, and the complexity of the support system. Mirror substrates that are more flexible require more precise support forces (either passive or active) to maintain the same optical precision.

Active optics apply slow correction forces to the mirrors (timescales of minutes or hours) in order to compensate for figure errors including those due to residual and transient errors in the support. These corrections serve to lessen the absolute requirements on the mirror support system. Active optics are also discussed in detail in [Chap. 5](#) and by Noethe (2002). We discuss active optics later in this chapter as they are critical to the mirror figuring process ([Sect. 3.3](#)) and to the performance in the telescopes ([Sect. 4.8.2](#)).

1.5.6 Low Thermal Inertia (Mirror Seeing)

Making the mirror substrates thinner and lighter provides another performance gain by lowering their thermal inertia. Ideally, the surface of the mirror (as well as the surrounding telescope) should be in α precise thermal equilibrium with the ambient mountain air. A thick mirror will tend to lag behind the temperature of the cooling mountain air on a typical night due to the thermal inertia of the glass. If the mirror (or the surrounding support structure) is out of temperature equilibrium, then mixing of the air by convection and/or wind can create pockets of warmer and colder air with higher and lower refractive index which blurs the image. To minimize this “mirror seeing,” the mirror surface should track the ambient temperature within about 0.2°C (Siegmond et al. 1990). This means the mirror substrate should have a thermal time constant less than 1 h in typical conditions. Such a short time constant can be achieved in the honeycomb structure by actively ventilating the inside of the mirror structure since the thickest section of glass is only 28 mm thick at the faceplate.

1.5.7 Low CTE and Fast Thermal Equilibrium

Any thermal difference within or across the mirror substrate will induce a distortion in the optical figure if the coefficient of thermal expansion α is nonzero. Local differences in temperature δT through the body of the mirror can produce bumps (or dips) on the surface. These bumps are proportional to the CTE, the temperature variation, and the thickness h ($\alpha\delta T h$) (Angel et al. 1983). Vertical temperature gradients also cause bending proportional to $(\alpha\delta T/h)$. This thermal distortion has been an issue since telescopes were first invented. An approach in the mid-twentieth century has been to make telescope optics from newly invented materials with effectively zero CTE at room temperature (Cervit, Zerodur, ULE). Unfortunately, these materials are only available in the form of solid slabs or meniscii and are not amenable to forming large light weight structures. The zero-CTE solutions are optimal for thermal distortion, but they are generally poor for mirror seeing because their increased mass and thickness makes them frequently out of thermal equilibrium. An alternative approach is to use a material with high thermal conductivity to suppress temperature gradients. A thermally optimized substrate represents a trade-off of short thermal time constant, high thermal conductivity, and low CTE (Miroshnikov et al. 1992).

Variations in CTE ($\delta\alpha$), like variations in temperature, induce thermal distortion in the mirror surface. The polished mirror must be stable against temperature changes since it will work in a mountaintop environment that may be 40°C colder than the optical laboratory. This requires the mirror substrate to have a homogeneous CTE so that uniform temperature changes (ΔT) do not induce distortions in the surface ($\delta\alpha\Delta T$) (Angel et al. 1983).

1.5.8 Affordable

The cost of a telescope mirror is perhaps the most difficult parameter to specify from first principles, although there is broad agreement that the cost should be optimized downward. Perhaps we can conclude that we must contain the mirror budget within a reasonable fraction of the total telescope project. Elaborate fabrication schemes such as machining an 8-m honeycomb blank out of zero-expansion glass ceramic have been ruled out for that reason. The same machined honeycomb technology may well be practical for a space telescope where the budget-to-aperture ratio is significantly higher. Cost may also be a significant factor when considering the support and handling schemes for large mirrors.

1.5.9 Compact Focal Lengths

Independent of the substrate technology, keeping the telescope optics compact may significantly impact the total cost of a telescope and its enclosure. The pure construction costs have to be traded with the tighter alignment tolerances of fast optics. Meinel (1979) discussed how the costs of larger telescopes might be reduced below the traditional $D^{2.7}$ scaling law. Factors that significantly affect the scaling are the focal ratio of the optics, the structural configuration of the telescope, and the configuration of multiple or segmented apertures. Telescopes with primary mirrors approaching $f/1$ (focal length equal to the diameter) seem to be optimal in controlling the overall cost of the telescope and enclosure, and have been demonstrated to fall below the traditional scaling relation. The technology for fabrication and testing of fast paraboloidal surfaces near $f/1$ is discussed below. LBT (↪ Sect. 4.8) is an example of a large telescope that included short focal ratios, multiple primary apertures, and novel structural design in its construction.

2 Casting Borosilicate Honeycomb Mirror Blanks

2.1 Key Concepts

The borosilicate honeycomb mirror blanks developed by the Steward Observatory Mirror Lab at the University of Arizona are continuous glass structures with a sandwich geometry. By melting chunks of glass into a mold, two facesheets are formed with a hexagonal rib structure in between to provide mechanical stiffness. The key concepts involved in that development are the following:

1. Complex shapes can be formed by remelting borosilicate glass into a complex ceramic mold that is the negative of the desired shape. A version of this process was used by Corning in

1935 to produce the blank for the Hale 5-m mirror with 100-mm wide ribs on the back side of the mirror to improve stiffness. SOML has refined the process to achieve an increased light weighting factor ($\times 5$) with 12-mm wide ribs and increased stiffness with a nearly continuous backsheet. Details of the molding process are discussed below in [Sect. 2.2](#).

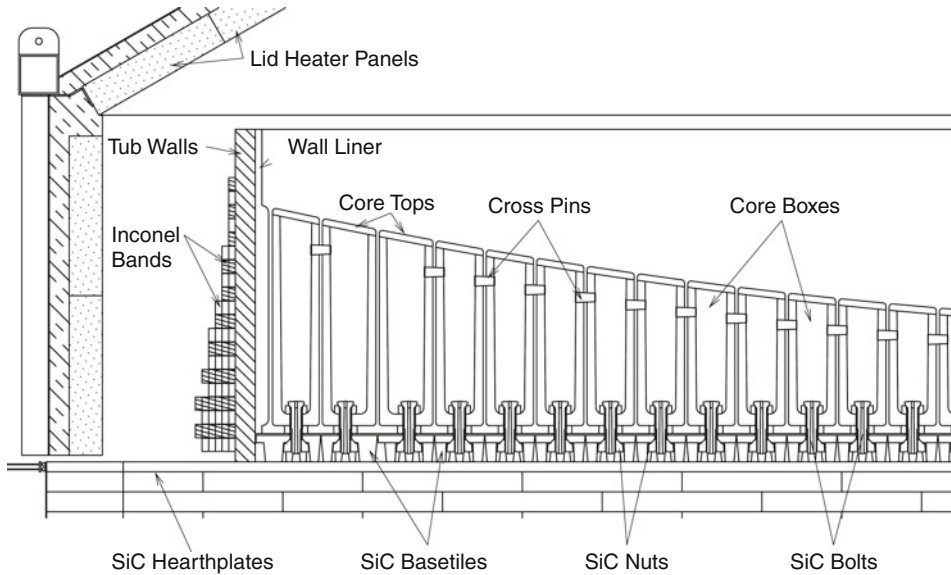
2. The furnace and mold are spun during the remelting process to produce the rough paraboloidal shape by centrifugal force. The radius of curvature of the molten glass surface is controlled by the rotation rate. When a liquid is rotated with angular velocity ω in a gravity field, the surface takes on the shape of a paraboloid with radius of curvature $R = g/\omega^2$, where g is the acceleration due to gravity. In addition to forming the desired surface shape, the spinning eliminates the need for an enormous amount of rough grinding. The glass that would need to be ground away (without the spinning) is greater than the mass in the remaining honeycomb sandwich structure for the fast paraboloidal surfaces being produced.
3. Existing industrial technologies are used for furnace materials and controls. The mirror blanks are formed in a rotating electrical furnace with heated surfaces all around the perimeter, base and top. A crossed roller bearing allows the entire furnace to rotate while slip rings bring the electrical power onto the rotating platform. While this particular application is novel, all of the furnace technologies are in common use for steel production and similar industries. Additional details on the furnace and the casting facility are discussed below.
4. After remelting, the resulting glass structure must be cooled and annealed carefully in order to achieve a low-stress blank and the long-term stability and safety which that entails. This slow cooling is the most challenging part of the process from the controls point of view as precise temperature control is needed for several months.
5. After the glass reaches room temperature, it is necessary to remove the mold material from the inside of the glass honeycomb to produce a useable light weight mirror blank. The aluminosilicate mold material must withstand chemical attack from the molten glass, it must withstand the hydrostatic forces of the molten glass, yet it must also be friable enough to be removed from inside the glass structure. The mold material is mechanically broken up using high pressure water spray which erodes the fibrous ceramic mold without damaging the glass surface. The surface of the glass has a texture matching that of the mold.

2.2 Process Description

2.2.1 Mold Assembly

The casting of one of three 3.5-m diameter blanks has been described in detail by Goble et al. (1989). The procedures used for mold assembly and casting have been described by Hill and Angel (1992) and Olbert et al. (1994) for the first of five 6.5-m mirrors. This section describes the mold assembly process for an 8.4-m honeycomb blank. The two 8.4-m primaries for LBT have been described by Hill et al. (1998) and Martin et al. (2006).

A large cylindrical tub is assembled to contain the entire mold under the hydrostatic pressure of the glass that will flow into it. The viscosity of the glass is similar to that of honey on a cold day (10^3 P) and is sufficiently low that the full hydrostatic forces are applied to the mold. The base of the mold is assembled from a close-packed array of individual silicon carbide (SiC) basetiles aligned with the positions of the honeycomb core molds that will form the voids in the honeycomb. Each hexagonal basetile has a captive SiC nut below it which holds the SiC bolt



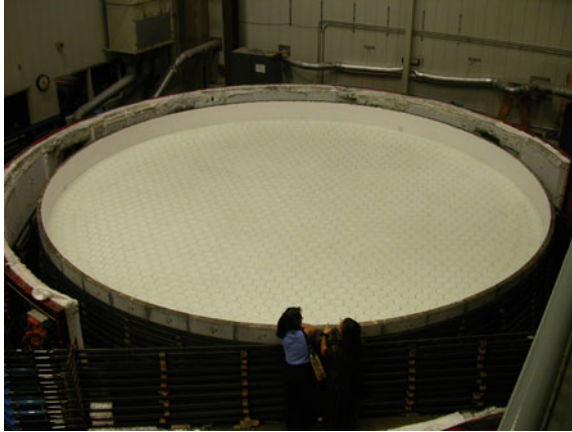
■ Fig. 4-1

This figure shows a cross section of the mold for the 8.4-m diameter honeycomb mirror. Glass chunks are piled directly on top of the cores before the casting process begins. The furnace boundary is shown on the left side. Steward Observatory drawing by Eric Anderson

used to restrain the core mold from floating under buoyancy forces. The basetiles can be seen attached to the bottom of the mirror mold in ► Fig. 4-1.

The classic failure mode of such a mold is having the core boxes float in the molten glass. This is a challenging problem because common fastener materials such as steel melt at or below the casting temperature. The reusable injection molded SiC bolts and nuts used to hold down the core boxes are manufactured by Ferro Corporation. Inner and outer tub wall sections are assembled around the tile base in the manner of barrel staves. The segments of the tub wall are made from Carborundum *Carbofrax* SiC-based castable refractory. The barrel-like walls are restrained by bands of *Inconel 601* steel which is the last alloy that will survive in an oxygen atmosphere at these temperatures. Each band wraps 90° around the tub and connects to pneumatic cylinders outside the furnace in order to take up slack and maintain tension as the band expands with temperature. The pneumatic cylinders constrain the tub against hydrostatic pressure and centrifugal forces during casting but can be relaxed to avoid stressing the honeycomb blank during cooling. The tub walls and the restraining bands can be seen in ► Fig. 4-2. The 8.4-m mold uses a total of 124 *Inconel* bands with a cross section of 12.9 square cm each.

The tub and basetiles are lined with aluminosilicate refractory fiberboard to avoid chemical reactions between the SiC and the molten glass. Inside the tub, the 8.4-m LBT mold has 1,662 ceramic fiber core boxes which form the voids in the honeycomb structure. Each of these core boxes is manufactured by Rex Roto Corporation and then machined to its final shape at the Mirror Lab. Olbert and Schenck (1997) report on several studies of the material properties of this particular ceramic fiber and rigidizer formulation. The friable aluminosilicate fiberboard has reasonable structural strength and is only mildly attacked by the chemistry of the molten



■ Fig. 4-2

This photo shows the 8.4-m GMT1 primary mirror on the furnace hearth as it was being de-molded after casting in October 2005. The SiC refractory tub walls surrounded by the Inconel steel bands which constrain the hydrostatic pressure of the molten glass can be seen. One of eight sets of pneumatic cylinders which tension the bands can be seen in the lower left corner. Heater elements in the walls of the furnace can be seen near the top. The honeycomb core molds can be seen through the partially transparent surface of the glass. Because this mirror blank will be an off-axis section of a larger parent paraboloid, no central hole is required



glass. Each machined core box is aligned with its particular hex tile using a special box placement tool in order to maintain the uniformity of the honeycomb ribs. After the SiC bolt has been tightened to hold each core to the bottom of the mold, stabilizing cross pins are installed to prevent the cores from leaning to the side during the casting and spinning. Post collars separate the body of the core from the floorboard to form the holes in the perforated backplate of the mirror. Finally, a lid is glued and pinned onto the top of each core box. Before the mold is loaded with glass, it is heated to the glass melting temperature to cure all the ceramic glue and to stabilize the shrinkage.

2.2.2 Furnace Description

The process for spincasting large borosilicate honeycomb mirrors requires heating tons of glass in a complex mold to $1,165^{\circ}\text{C}$ while spinning the entire furnace at 6.8 rpm. The peak power consumption during heating exceeds one megawatt. After casting, the honeycomb blank must be cooled through the annealing temperature range at $0.1^{\circ}\text{C}/\text{h}$. The glass is in the furnace on a controlled temperature schedule for a 14 week period. The furnace control system reads 600 N-type thermocouples and controls 270 8-kW heaters. The control system regulates the furnace temperature to a few degrees over the entire casting cycle. Considerable design effort has gone into assuring that a component failure or a control system error does not turn an 8-m mirror into an expensive patio ornament. Errors are avoided by four strategic steps: fault avoidance, fault detection, fault containment, and fault recovery. System redundancy begins with three onboard 68,000-family VME-bus computers (powerful when they were new) which control overlapping

areas of the furnace. Redundancy extends down through the temperature measurement and power control systems with many modular, interleaved, and optically isolated subsystems. Data logging and system monitoring are achieved with a pair of Sun workstations running IRAF software in the control room. Rotation of the furnace is controlled by two 40-horsepower DC servomotors with speed regulation to 0.1%. A 6-m diameter furnace was constructed on a larger turntable in 1986 to cast three 3.5-m blanks. The furnace was expanded in 1990 on the same turntable to cast 6.5- and 8.4-m blanks. The full-size furnace has an internal diameter of 9.5 m. There is a cylindrical wall 1.2 m high, topped by a flat disk and a conical lid section. The furnace has as major components a hearth, walls, and lid. The circular bottom plate, or hearth, is pieced together from smaller pie-shaped segments of silicon carbide-based refractory. The hearth provides the working surface on which to build the complex honeycomb mold described above. The hearth, walls, and lid contain embedded electric heater elements to heat the furnace. Additional details on the furnace and its control system are provided by Hill et al. (1990).

2.2.3 Casting

For each 8.4-m LBT mirror, 19 metric tons of glass were carefully loaded onto the mold and the furnace lid was lowered into place. The chunks of glass must be placed carefully by hand in order to avoid damage to the fragile ceramic mold material as shown in  Fig. 4-3. Heating of the glass and mold begins after the top is placed on the furnace. The temperature is ramped up to the maximum temperature over the course of a week. The rate of heating is controlled in order to limit the thermal stresses on the refractory mold components. Rotation at 6.8 rpm begins when the furnace reaches 750°C – the temperature at which the glass blocks just begin to flow. Spinning and heating continue until the maximum temperature of 1,165°C is reached. The glass blocks slump and flow together to form a meniscus on top of the honeycomb mold. Then the glass slowly flows down the gaps for the honeycomb ribs as the viscosity decreases over several hours. The furnace temperature is held at the maximum temperature for 5 h to allow the glass to flow into all corners of the mold and to allow trapped air bubbles to rise to the surface. This process is monitored by CCD cameras looking through pinholes in the furnace wall. Flash lamps are used to provide blue light in order to achieve contrast against the strong red blackbody radiation. A sequence of these images showing the glass blocks slumping is shown in  Fig. 4-4. The temperature is ramped back down to 650°C over the next 2 days while the furnace continues to spin. At this point, the glass is viscous enough to hold the paraboloidal shape without spinning, but is still soft enough to flow microscopically under mechanical or thermal stresses.

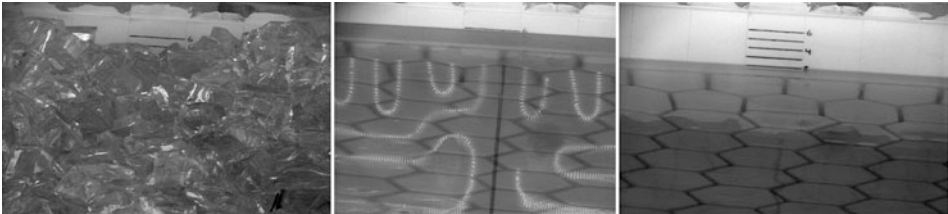
2.2.4 Annealing and Cooling

In order to produce a low-stress glass blank at the end of the process, the glass must be cooled slowly through the annealing range. The annealing range is that range of temperatures where the relaxation time of stresses in the glass increases from hours to years. Any stresses leftover from the casting process relax quickly away above the annealing point (upper end of the annealing range). The blank must be cooled very slowly through the annealing range in order to minimize the thermal gradient (due to the cooling) that becomes a stress gradient after the blank reaches



■ Fig. 4-3

This photo shows the loading of the 19 tons of Ohara E6 borosilicate glass chunks on *top* of the honeycomb mold for the casting of the LBT2 8.4-m primary in June 2000. The mold is assembled inside the rotating furnace at the Steward Observatory Mirror Laboratory. Each 5 kg chunk of pre-inspected glass is carefully placed by hand in order to protect the fragile mold material



■ Fig. 4-4

These photos show the melting sequence of the E6 glass during the casting of the LSST primary mirror in April 2008. The *left photo* shows the glass as discrete chunks while the furnace is heating at a temperature of 300°C. The *middle photo* shows the glass chunks merged into a meniscus on *top* of the honeycomb mold at a temperature of 966°C. The zigzag lines are the reflections of electric heater elements in the lid of the furnace. The *right photo* shows the finished product with the glass flowed into the honeycomb mold while the blank was annealing. The horizontal lines show the depth of the faceplate glass in inches. The outlines of the core molds can be seen through the transparent glass surface

thermal equilibrium at room temperature. Annealing theory is discussed by Morey (1938). The annealing time increases as the square of the thickness for a solid slab of glass. Because the mold inside the honeycomb structure is more conductive than glass, the annealing time increases roughly linearly with the thickness of the honeycomb sandwich. In practice, the mirror is cooled at a rate of $0.1^{\circ}\text{C}/\text{h}$ in order to achieve a thermal gradient less than 1°C through the thickness of the blank. This cooling rate is maintained from 530°C to 450°C which is a conservative estimate of the annealing range. The entire surface of the furnace around the mold is controlled to a fraction of a degree in order to make the entire mold nearly isothermal. The cooling rate increases below the annealing range, but care must be taken to avoid thermal stresses in the now rigid glass structure. Thermal gradients of 10°C across the blank would induce dangerous stresses. The honeycomb sandwich is much more sensitive to thermal gradients than borosilicate bakeware because of its rigid structure. The total casting process uses 9 days for the initial heating and spinning, plus an additional 90 days for annealing and cooling back to room temperature.

2.2.5 Mold Removal and Handling

A handling fixture composed of a 10-m diameter lifting frame, and turning ring is used to lift, support, and flip (invert) blanks up to 8.4-m diameter as they are removed from the furnace. The lifting strategy is to insure that tensile stresses in the glass blank stay less than 0.7 mPa at every point in handling. The first lift of the mirror off the furnace and into a vertical position causes the largest stress it is ever likely to see because the honeycomb structure is carrying the extra 16-ton load of the refractory mold material. Some finite element analysis of the handling support is described by Parodi et al. (1992). The designs of the support systems for holding the 8.4-m mirrors in the furnace, in the handling fixture, and in the telescope are summarized by Parodi et al. (1997). Additional details about the lifting and handling of these large honeycomb mirrors are provided by Davison et al. (1998).

In the first step of lifting, the cast mirror off of the furnace hearth, 36 mild steel pads 60 cm in diameter are positioned and attached to the blank's faceplate with a 1-cm thick patchwork layer of RTV silicone sealant and silicone rubber tubing. The silicone tubing acts as a portal for easy removal of reaction products released during cure, letting the RTV silicone cure rapidly and uniformly throughout the layer. Shear strength, tensile strength, and creep response of steel-RTV-glass sandwiches are measured to assess the short- and long-term performance of the attachment method. The glass blank plus mold material are lifted off the furnace hearth with the steel lifting frame attached to these 36 pads. Rubber spacers between the frame and the pads assure that the load is spread uniformly across the glass blank using all of the pads. The lifting frame with the glass blank attached is mounted horizontally in the turning ring. To clean the mold material from the blank, the turning ring is rotated until vertical. This exposes the backplate of the mirror which had been resting on the hearth. A special washout stand with two independent elevating work platforms is assembled and moved within easy reaching distance of the backplate. Working from top to bottom, the cleaning crew removes all the SiC nuts and hexagonal tiles which are the reusable parts of a mold. Then working from bottom to top, each cell box is breached and the SiC bolt and washer removed. Using a high pressure (11 mPa) water jet "cleaning wand" which breaks up the friable ceramic, the core material is sliced into chunks and the pieces of soft refractory mold are removed by hand. This particular order of operation is done to avoid filling the mold with additional tons of water. Once the mold

material has been removed from the glass honeycomb, the blank is rotated again with the same lifting fixture to a downward facing position to prepare for machining the rear surface of the blank.

2.3 Why E6 Borosilicate Glass?

The honeycomb sandwich blanks produced at the Steward Observatory Mirror Lab are all made of remelted E6 glass supplied by Ohara Incorporated of Kanagawa, Japan. E6 is a borosilicate crown glass produced in 1-ton batches by traditional glassmaking techniques in large clay pots. Glass is used as the substrate for a great majority of optics because its amorphous structure can be polished very smooth without problems with crystal grain structure in the surface. Glasses hold their shapes indefinitely as long as they are not stressed to the point of fracture. Contrary to urban legend, the glass in very old windows or telescope mirrors does not flow (Zanotto 1998). Thus, a precise figure produced in the optical shop can be stable for centuries in the telescope. Moreover, glass is perfectly elastic, meaning a shape change due to applied stress disappears completely when the stress is removed, provided no fractures are induced. Metals and other candidate materials may deform inelastically under applied loads. The optical figure produced in the shop, with the glass mirror on a given set of support forces, will be reproduced in the telescope when the mirror is supported with the same forces. A limitation of glass mirrors is that stresses must be kept low enough to avoid fractures for the life of the telescope. The maximum allowed stress depends on the area and processing of the glass surfaces but is generally on the order of several MPa.

Borosilicate glass was chosen for this application because it can be readily worked and formed at modest temperatures below 1,200°C. Other glasses such as fused silica (SiO_2) must be heated to near their decomposition temperature before they can be slumped or molded. Typical glass ceramics cannot be remelted after the initial production. The ideal mirror substrate would have a zero coefficient of thermal expansion (CTE). Borosilicate glass has a reasonably low CTE and excellent chemical durability. In particular, E6 has a room temperature CTE of $27.8 \times 10^{-7}/^\circ\text{C}$ which is about 15% lower than typical borosilicate glasses. For ground-based telescopes, one can compensate for the finite CTE by ventilating the honeycomb structure to follow the ambient air temperature on the mountain. E6 is supplied in 5-kg chunks with broken surfaces as it is mined out of the 1-ton batch pots. These chunks are particularly attractive for the remelting process as the broken glass surfaces are uncontaminated and fuse back together without any detectable bond line in the final honeycomb structure. Each individual chunk is inspected under polarized light to permit any impurities to be removed before the remelting process.

Empirically, the batch production of E6 allows very careful control of the chemical composition and therefore the homogeneity of CTE. The variation of CTE from batch to batch (1-ton batches) is $0.6 \times 10^{-7}/^\circ\text{C}$ p-v for production over several years. The Mirror Lab typically sorts the batches by measured CTE to get a range of only $0.4 \times 10^{-7}/^\circ\text{C}$ within a single mirror. The glass chunks are further mixed as they are placed in the mold to average any residual batch-to-batch CTE variations across the blank. This low variation of the CTE is important as it allows the mirror figure to be stable against the seasonal temperature changes on the mountaintop. The mirror is able to retain the figure that it had in the optical laboratory as the ambient temperature cools by 40°C.

2.4 Blank Dimensions and Properties

2.4.1 Honeycomb Sandwich Design

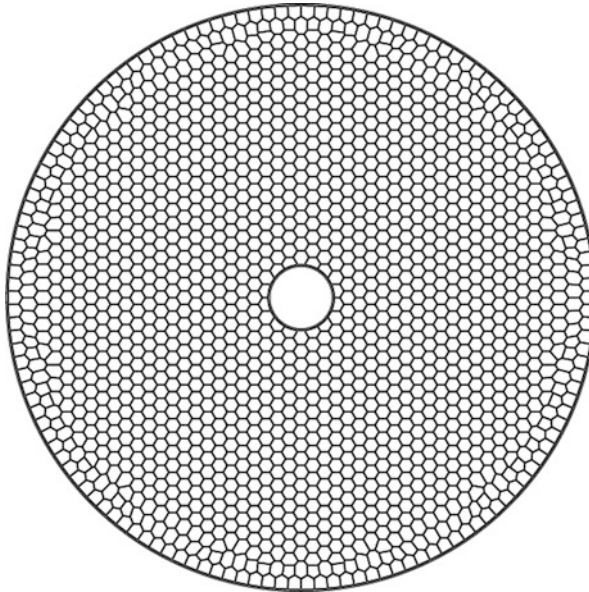
The basic concept of a honeycomb sandwich structure is to have a light weight core separating two thin plates. This type of structure is encountered commonly in everyday life with examples as varied as corrugated cardboard boxes or airplane wings. Often, the central core structure is composed of ribs as the two-dimensional analog of an I-beam. The bending stiffness comes mainly from the depth of the sandwich between the two faceplates. The light weight core can have any geometry ranging from foam to posts to ribs. Structures with equal mass in the facesheets and core seem to be the most mass efficient. The shear stiffness comes from the structure of the core, so a pattern of interlocking ribs is quite efficient. For a telescope mirror, the internal structure should have a uniform thermal time constant and therefore a uniform mass density such as a regular pattern of square or hexagonal ribs. Hexagonal ribs resist the hydrostatic pressure of the glass best because they have the shortest length for a given area of the core. Fillets are added where glass plate structures intersect to avoid stress concentrations at sharp corners in the structure.

2.4.2 Dimensions

As described by Hill et al. (1998), the finished outside diameter of the honeycomb mirror blanks for LBT (◆ Sect. 4.8) is 8.417 m with an optical aperture covering 8.405 m. This is the largest diameter blank that SOML can produce for a number of practical reasons relating to the sizes of doors, furnace, and handling fixtures. The central hole inside diameter is 0.889 m with an optical aperture of 0.901 m. The finished facesheet thickness is 28 mm. The facesheet thickness is set at 28 mm in order to reduce the thermal time constant below 1 h (with forced-air ventilation). Given that thickness, the 192-mm spacing between ribs is chosen to limit deflection of the unsupported facesheet under gravity and polishing pressure. Both the faceplate diameter and thickness are cast with a 10-mm thickness margin to be removed during generation. The finished backplate thickness is 25 mm, and the ribs are cast to a thickness of 12 mm. The outer edge thickness of the blank is 894 mm, while the inner edge thickness is only 437 mm. The 457-mm sag in the paraboloidal surface is a result of the 9.600-m focal length ($f/1.142$). The pattern of the honeycomb ribs is shown in ◆ Fig. 4-5. Unlike previous smaller mirrors, the transition from the regular hexagon pattern to the circular edge of the mirror occurs over three core widths. Each of the irregular core shapes at the edge is optimized for minimum stress of the core bottom under the buoyant force. Having approximately uniform sizes of these irregular cores also aids in the fabrication of the mold. The finished weight of the mirror blank is just over 16 metric tons. The honeycomb structure is slightly over 20% of the density of solid glass.

3 Optical Fabrication and Testing of Large, Aspheric Mirrors

While the casting process is unique to borosilicate honeycomb mirrors, the rest of the manufacturing process could be applied to any large telescope mirror in most respects. A number of new manufacturing techniques have been developed for honeycomb mirrors, largely because



■ Fig. 4-5

This figure shows the pattern of honeycomb ribs for the 8.4-m mirror. Each of the ribs is 12 mm thick. One thousand six hundred sixty-two core boxes make up the honeycomb structure. A 28-mm thick faceplate and a 25-mm thick backplate complete the honeycomb sandwich structure. Steward Observatory drawing

the same telescope projects that use honeycomb mirrors have chosen to use relatively fast primary mirrors, $f/1.25$ for the MMT and Magellan telescopes and $f/1.14$ for the LBT. The segmented GMT primary mirror is $f/0.7$. The techniques developed for honeycomb mirrors are therefore tuned to very aspheric surfaces. This has led to new technology for both polishing and measurement. In addition to the emphasis on asphericity, the manufacturing techniques that have been developed are appropriate for stiff mirrors, make use of the same support system that is optimized for telescope performance, and incorporate thermal control. Additional new technology has been developed to measure the off-axis segments of the GMT.

3.1 Overview of Fabrication and Testing

Following the production of a honeycomb mirror blank, the fabrication process can be divided into three stages in which the surface accuracy is gradually improved until it meets the specifications for the telescope. The first stage is fixed-abrasive grinding, also known as generating or machining. The cutting tool is a rapidly rotating wheel with embedded diamond particles. It is mounted on a computer-controlled mill and moved across the surface, removing glass to a controlled depth. The surface accuracy depends on the accuracy with which the tool position is controlled and on the accuracy of the measurements that guide the process. The surface figure is taken from an initial accuracy of about 0.5 mm rms (that of the cast surface) to around 10 μm rms.

The second and third stages are both lapping processes. The tool is a disk that is dragged over the mirror surface and wears away the glass. The tool's position in the direction normal to the surface is not controlled by the polishing machine; instead it rests on the surface with a controlled downward pressure. The rate of removal is generally assumed to follow Preston's relation (Preston 1927): it is proportional to pressure and relative lateral speed between the tool and the mirror surface. The surface accuracy is therefore limited by the predictability of removal rate and the accuracy of control of tool pressure, speed, and time spent on any part of the mirror.

The first of the two lapping stages is loose-abrasive grinding, in which hard abrasive particles roll between the mirror and a hard surface on the lapping tool. The surface accuracy is on the order of 1 μm rms and may be limited by the measurements available for the unpolished surface. This lapping also removes the subsurface damage (microscopic fractures) caused by the prior fixed-abrasive grinding.

The final stage is polishing, in which the hard surface of the lap is replaced by a softer surface such as pitch (the traditional material, a very viscous tar-like fluid) or synthetic polishing pads. Fine abrasive particles, generally one of several metal oxides, remove glass through a combination of chemical processes and mechanical abrasion. The surface becomes specular (reflective at visible wavelengths), microscopic roughness is reduced to about 1 nm rms, and subsurface damage due to loose-abrasive grinding is removed. The removal rate is low, on the order of 1 nm of depth for every meter of tool surface passing over a point on the mirror, and the surface can be controlled at the level of nanometers. While Preston's relation may be fairly accurate for the effects of pressure and speed, additional variables that describe the surface of the tool and the conditions of the abrasive appear to be important in polishing. The average pressure over the tool is generally well controlled, but the pressure distribution across the tool is more difficult to control, especially for very aspheric mirrors. The ultimate accuracy depends on the ability to control the pressure distribution and other variables that affect removal rate. Accurate measurements are critical. All of these aspects are discussed more in the following sections.

The preceding overview and the following subsections apply to the optical surface. The flat rear surface and edges of the mirror are also processed after the casting. The rear surface is machined flat, then lapped and polished to a specular finish. This provides an accurate interface for the mirror support system and allows visual inspection for flaws in the backplate and ribs. The lapping and polishing operations remove subsurface damage, leaving a surface that is essentially free of flaws and has maximum toughness against accidental impacts. After the rear surface is polished, the loadspreaders that will interface with the support system are surveyed into position and permanently bonded with a thin layer of compliant RTV adhesive. The mirror is then inverted to have the optical surface up. The inner and outer edges of the front facesheet and the backplate are also machined and lapped to provide accurate mechanical dimensions.

3.2 Accuracy Requirements

The general rule for surface accuracy of a ground-based telescope is that the optics must be more accurate than the best wavefront the atmosphere will deliver at all spatial scales. This rule holds whether or not adaptive optics is used. Without adaptive optics, the telescope's optics must not significantly degrade the images delivered by the atmosphere. With adaptive optics, most of the stroke of actuators in the deformable mirror should be reserved to correct the atmosphere rather than errors in the telescope optics.

The atmosphere induces large wavefront errors on large spatial scales but delivers a very smooth wavefront on small scales. The telescope optics therefore must be very smooth on small scales. The spectrum of wavefront errors after propagation through the atmosphere is usually described by a structure function, which gives the mean square difference in wavefront between points in the pupil as a function of their separation. The phrase “mean square difference” here refers to an average over all pairs of points with a given separation in the pupil. In the standard Kolmogorov model of turbulence, the structure function of the wavefront is

$$\delta^2(x) = \left(\frac{\lambda}{2\pi}\right)^2 6.88 \left(\frac{x}{r_0}\right)^{\frac{5}{3}},$$

where x is the separation in the pupil, λ is the wavelength, and r_0 is the coherence diameter or Fried’s parameter. The coherence diameter is related to the long-exposure image size θ (full width at half maximum) as

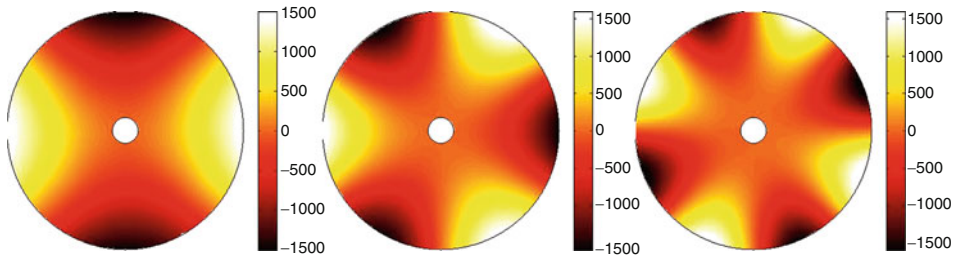
$$\theta = 0.98 \frac{\lambda}{r_0}.$$

Almost all of the projects using honeycomb mirrors have adopted a mirror figure specification of the same form as the atmospheric structure function. The one free parameter is r_0 or equivalently θ . A good telescope site will have median seeing of about 0.7 arcsec FWHM and exceptional seeing as good as 0.3 arcsec. A reasonable value of θ for the primary mirror figure specification is about 0.1 arcsec. In a perfect atmosphere, a mirror whose surface just meets this specification on all scales would form 0.1 arcsec images. In practice, it is easy to meet such a specification on large scales, so the finished mirror is much better than the specification on large scales while matching it on small scales. The actual image formed by the finished mirror (in a perfect atmosphere) has a nearly diffraction-limited core (width λ/D) and a halo of width θ .

The structure function for the atmosphere approaches zero at small separations. There is no need for the telescope optics to be perfect on small scales. The effect of small-scale errors whose rms wavefront error σ is a small fraction of a wavelength is to scatter a fraction of the light outside the localized point-spread function. This fraction, or loss, is $L = 1 - e^{-(2\pi\sigma/\lambda)^2} \approx (2\pi\sigma/\lambda)^2$. The mirror specification is relaxed on small scales to allow, typically, $L \leq 0.02$ at a wavelength of 500 nm.

The Kolmogorov structure function is dominated on large scales by wavefront tilt across the full pupil, equivalent to image motion, which changes slowly and is generally corrected by active guiding. For this reason, and because the mirror’s surface error contains no global tilt, the mirror specification is tightened on large scales to eliminate the effect of tilt induced by the atmosphere.

Because the mirror figure will be controlled with active optics in the telescope, there is a relaxation for low-order aberrations like astigmatism and trefoil. There is no point in trying to eliminate these aberrations by polishing because they will be adjusted in the telescope with small changes in the support forces to an accuracy that depends on the wavefront measurements at the telescope. More precisely, the figure errors that will be controlled with active optics are the low-order bending modes of the mirror. The bending modes form an orthogonal basis set for mirror figure errors and can be listed in order of decreasing flexibility defined as the ratio of rms surface distortion to rms correction force. (The rms correction force is the rms over all the support actuators.) The two most flexible bending modes look like the two components of astigmatism, and other low-order modes are also similar to low-order Zernike polynomials. For an 8.4-m honeycomb mirror supported by 160 actuators, each astigmatic mode has a flexibility of 130 nm



■ Fig. 4-6

Several low-order bending modes measured for the first LBT primary mirror. One each of the first three pairs of non-axisymmetric modes are shown. The rms force over the 160 actuators is, from left to right, 5 N for the astigmatic mode, 30 N for trefoil, and 85 N for quadrafoil. The color bars are labeled in nm of surface deflection

rms per N rms force. The average actuator's force at zenith is about 1,000 N, so several hundred nm of astigmatism can be corrected using a small fraction of the actuators' force ranges.

► *Figure 4-6* shows some low-order bending modes for the first LBT mirror (Martin et al. 2004). These modes were measured by bending the mirror with its active support system. The forces were calculated by analysis and applied to the mirror support system while the mirror was measured with the interferometer. Each plot is the difference between two measurements with different forces, so the mirror's figure errors cancel.

During manufacture, the mirror is supported passively, not on the active telescope support. The active optics correction is therefore simulated for figure measurements made in the lab. Correction forces are limited to a magnitude of around 30 N rms. Any low-order bending modes that are correctable within that constraint can be ignored during manufacture. The residual error after the simulated correction must meet the structure-function specification. Typically around ten low-order modes will be ignored in the lab, while up to about 30 modes might be corrected at the telescope.

Alignment of the mirror in the telescope provides additional degrees of freedom for active correction. The precise alignment is set according to wavefront measurements in the telescope, so the manufacturing requirement is to control alignment aberrations within the adjustment range at the telescope. For a Cassegrain or Gregorian telescope, the two axial degrees of freedom are the spacing between the primary and secondary, and between the secondary and focal plane. The corresponding wavefront aberrations are focus and spherical aberration. Most primary mirrors are conic sections of revolution with the surface height z , measured parallel to the optical axis, given as a function of distance r from the axis by

$$z = \frac{r^2}{R + \sqrt{R^2 - (1+k)r^2}}$$

R is the radius of curvature at the vertex and k is the conic constant. An error in R is equivalent to the focus aberration and an error in k is equivalent to spherical aberration. Tolerances for the LBT mirrors were 1 mm in R , equivalent to 6.9 μm rms focus in the surface, and 0.0001 in k , equivalent to 46 nm rms spherical aberration. Methods of measuring these alignment parameters are discussed in ► [Sect. 3.4](#).

The lateral position of the primary mirror in the telescope represents two more degrees of freedom (for x and y displacements). A displacement of the mirror's optical axis causes coma in the wavefront. The LBT specification allowed a 1-mm displacement between the optical axis and the mirror's mechanical axis defined as the line centered on the mirror's outer diameter and normal to the surface. This 1-mm displacement is equivalent to 620 nm rms coma in the surface. Measurement of centration is also discussed in [Sect. 3.4](#).

Active optics plays a broader role in the manufacture of the GMT off-axis segments. Three factors contribute to this difference. First, the seven segments must have the same focal length in order to maintain a common magnification over the 20-arcmin field of view. This means focus must be controlled much more tightly. Second, alignment aberrations are very different for an off-axis mirror. Translation in the off-axis direction – sliding the segment along the parent surface – causes primarily focus and astigmatism followed by coma. Rotating it within the parent surface causes astigmatism then coma. (For a symmetric mirror, only the coma due to translation is present.) Alignment in the telescope therefore provides two important degrees of freedom that can be used to correct figure errors due to imperfect manufacture. [Figure 4-7](#) shows that displacements of 1 mm can produce large changes in focus, astigmatism and coma.

The third factor making active optics so important for manufacture of the GMT segments is the difficulty of measuring low-order aberrations in the lab for a large off-axis mirror. The telescope will provide much better wavefront measurements using the optical system's natural geometry that brings collimated light to a focus, but this geometry is not practical in the lab. [Section 3.4](#) describes the optical test of the GMT segments. Small misalignments in the test optics, at the limit of the achievable accuracy, cause measurement errors of about 1 μm rms surface in focus and several hundred nm in astigmatism. These aberrations exceed the specification for figure errors at the telescope and are close to the amounts that can be corrected by bending the mirror with acceptable forces and residual errors. With a combination of alignment and bending, however, they can be corrected without exceeding the budget for either. One of the principles for manufacture of the GMT segments is that all low-order aberrations must be

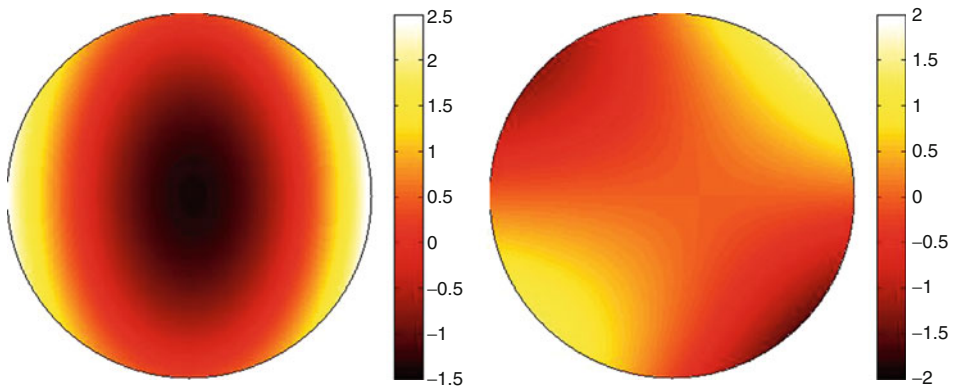


Fig. 4-7

Mirror surface change equivalent to displacements of the off-axis GMT segment along the parent surface. *Left*: a 1 mm shift in off-axis distance. *Right*: a 25 arcsec rotation of the segment about its center (1 mm across the 8.4-m diameter). The vertex of the parent is to the left. Color bars are labeled in microns of surface

controlled to an accuracy such that they can be eliminated with a combination of alignment and bending once they are measured accurately at the telescope.

3.3 Polishing of Large Mirrors

The challenge of polishing telescope mirrors comes largely from their asphericity. Their size implies a need for some large and expensive equipment but does not constitute a fundamental challenge. Polishing a spherical surface makes use of a powerful principle known as passive smoothing: as the tool moves across the mirror surface, the surfaces of the tool and mirror naturally come to a common shape with constant curvature, with little or no explicit guidance by the optician. A bump on the mirror surface automatically sees higher pressure from the tool and is smoothed down to the correct height. Any deviation from this principle makes it more difficult to achieve a smooth and accurate surface. The honeycomb mirrors have faster focal ratios and are therefore more aspheric than other telescope mirrors. The LBT primary mirrors have 1.4-mm peak-to-valley departure from the closest spherical surface, a large departure for optics that need to be accurate to a few hundred nm on large scales and the order of 10 nm on small scales. The GMT off-axis segments have 14-mm p-v departure. The difference is not as severe as it appears because the asphericity of the off-axis mirrors is largely astigmatism, with gentler curvature changes than the spherical aberration of a symmetric mirror.

Many methods of dealing with asphericity have been developed over centuries and especially in recent decades. The problem with a traditional lap, or polishing disk, is that it can match the shape of the surface at one position but not at other positions. The misfit causes large pressure variations across the contact surface, leading to uneven and unpredictable local removal of glass. These problems can be reduced by using small tools (so the magnitude of misfit is limited to a few μm) or flexible tools that sag to match the mirror surface. They can be eliminated by using figuring process such as ion beam figuring (Wilson and McNeil 1987) and magneto-rheological finishing (Prokhorov et al. 1992; Jacobs et al. 1995) that remove material without abrasion with a lap. These methods have been successful in a number of applications. They have the limitation that they give up or compromise the principle of passive smoothing that automatically creates smooth spherical surfaces.

Nelson and colleagues (Lubliner and Nelson 1980) preserved the smoothing principle when they developed stressed-mirror polishing for the aspheric segments of the Keck telescopes. Forces and moments are applied to the edge of the mirror to bend the desired aspheric surface into a sphere, which is then polished with a large, stiff lap. Release of the applied stresses lets the mirror relax to the desired shape. The method is powerful but not accurate enough to produce a finished mirror in part because it cannot easily be applied to segments with a polygonal outline, so the Keck segments and similar segments have been finished with methods such as those listed in the previous paragraph (Allen et al. 1992).

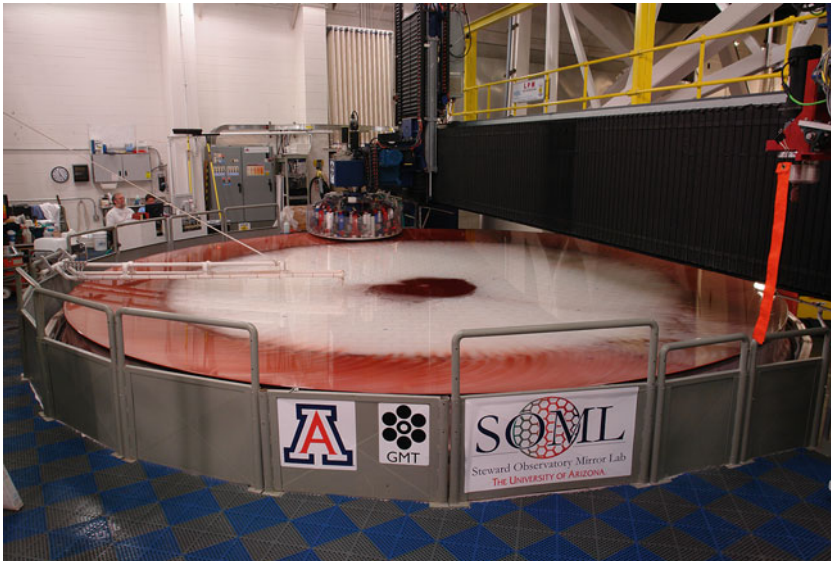
The Steward Observatory Mirror Lab developed stressed-lap polishing, which shares some of the principles of stressed-mirror polishing but bends the lap instead of the mirror (Angel 1984; Martin et al. 1988). This method is most appropriate for large, stiff mirrors. (It would be impossible to bend an LBT mirror into a spherical shape without breaking it.) It allows use of a relatively large, stiff lap that has strong passive smoothing on scales up to a large fraction of a meter. Control of the surface on larger scales is achieved by figuring, that is, by varying the pressure, speed, and dwell time as a function of position on the mirror. A stressed lap bends dynamically as it moves over the mirror surface so as to match the local shape at any point. A set

of actuators on the circumference of the polishing disk apply bending and twisting moments to induce low-order shape changes.

This principle can be applied to any aspheric surface. One can think of the stressed lap as matching a Taylor expansion of the mirror surface. A rigid passive lap resting on the surface matches a low-order expansion consisting of a constant and 2 tilts. By adding curvature variations, the stressed lap matches a high-order expansion of the mirror surface including curvature in orthogonal directions and several higher derivatives. For a given level of misfit, a stressed lap can be many times larger than a passive lap and, therefore, it gives passive smoothing over a larger range of scales.

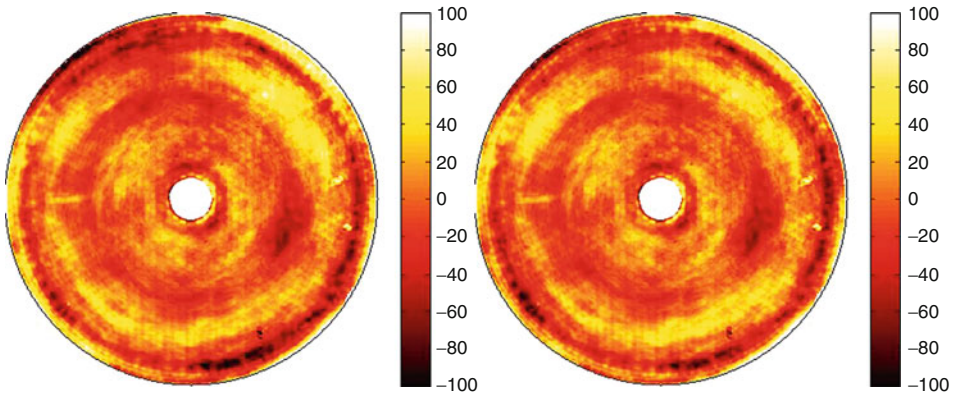
Stressed laps up to 1.2-m diameter have been used for honeycomb primary mirrors (► [Fig. 4-8](#)) and down to 30 cm for smaller secondary mirrors. Stressed-lap polishing has generally been supported by a complementary process that uses small, passive laps to bridge the gap between the scales of strong passive smoothing and well-controlled figuring. One such process is a so-called rigid-conformal lap of 10–30 cm diameter used with a small orbital motion (Kim and Burge 2010). Varying its dwell time as a function of position on the mirror gives highly predictable figure changes.

► [Figure 4-9](#) and ► [4-10](#) show color maps of the surface error in the finished LBT primary mirrors as measured in the Mirror Lab. The mirrors had been installed in their operational support cells, and active optics were used to control the first 43 bending modes. (Correction forces were less than 100 N.)



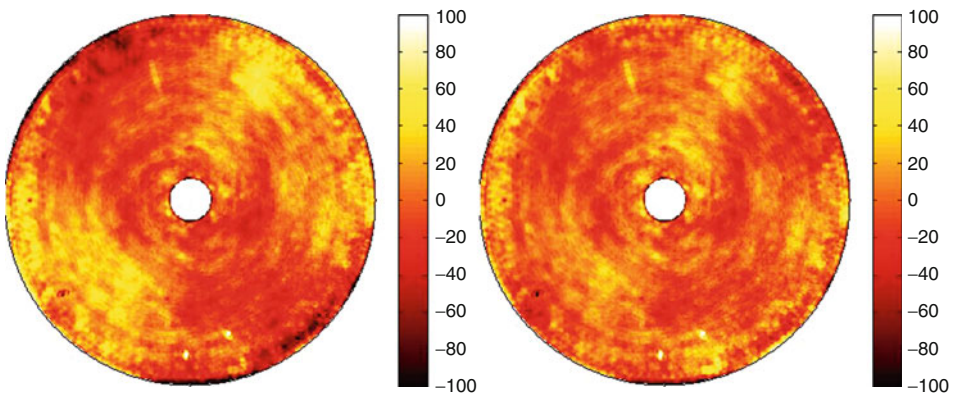
■ Fig. 4-8

Polishing with the stressed lap for the GMT1 primary mirror in 2009. The 18 actuators bend the lap continuously to match the local curvature of the mirror surface. The plastic pipes on the left distribute polishing slurry (red) across the mirror surface (Steward Observatory photo by Ray Bertram)



■ Fig. 4-9

Measured surface error for the first 8.4-m LBT primary mirror. The color bar labels surface error in nm. At *left* is the direct measurement with the mirror on its active support, with an accuracy of 28 nm rms. At *right*, slight residual astigmatism and spherical aberration are subtracted, giving 27 nm rms



■ Fig. 4-10

Same as [Fig. 4-9](#), for the second LBT primary mirror. The direct measurement gives an accuracy of 25 nm rms. Subtracting astigmatism and spherical aberration gives 18 nm rms

3.4 Measurement of Large Mirrors

The basic method that gives the most accurate measurements of telescope optics is phase-measuring interferometry, in which the full mirror surface is illuminated with a coherent wavefront (the test wavefront), and the reflected wavefront is combined with an accurate reference wavefront from the same laser. This produces an image of the mirror overlaid with interference fringes and ultimately a contour map of the mirror surface with a resolution of $\lambda/100$ or better. In this method, the test wavefront serves as a template that the mirror surface is compared with, so the test wavefront must be shaped to match the desired surface. This is done with a set of optics called a null corrector, which can be as simple as a single lens for mildly

aspheric mirrors or a combination of lenses, mirrors, and/or computer-generated holograms for more severe aspheres. The mirror is made to match the template wavefront, so the accuracy of the null corrector is critical.

The interferometric test requires normal incidence to make the reflected light return to the interferometer, so the light source and null corrector are located near the mirror's center of curvature, that is, a distance R away from the mirror. Even the fast LBT primary mirrors have $R = 19.2$ m, so a large test tower is required to support the test optics and hold them stable relative to the mirror.

The null corrector for an LBT mirror converts a spherical wavefront into one that will match the 1.4-mm asphericity of the mirror after it propagates 19 m to the mirror surface. The null corrector consists of two large lenses, 200–300 mm in diameter. All axial dimensions, including the spacing between the lenses and their separation from the interferometer, must be controlled with challenging tolerances on the order of 10 μm . It is difficult to rule out the possibility of a mistake such as the one that occurred with the primary mirror for the Hubble Space Telescope (HST). Because the LBT mirror is 30 times more aspheric than the HST primary, even a much smaller mistake could not be tolerated.

Burge (1993a, b) developed an independent test of null correctors that has been used for all the large honeycomb primary mirrors (except for the GMT segments discussed below). It consists of a small computer-generated hologram designed to mimic a perfect telescope mirror. Placed close to the null corrector, this certifying hologram returns light to the interferometer along the same ray paths as the light that would be reflected from a perfect primary mirror. The return wavefront is produced by diffraction from a pattern of rings written on a glass substrate. The hologram's design and manufacture are independent of the null corrector, apart from the fact that both are based on the prescription of the mirror. This method has been successful in validating the measurements of primary mirrors for LBT, MMT, Magellan, and other telescopes using honeycomb mirrors.

In routine use of the interferometric test, the mirror is aligned to the interferometer and null corrector (or vice versa) to minimize the measured aberrations. This allows the template wavefront to shift relative to the physical mirror surface to find the best match. In order to control radius of curvature and centration of the optical axis, the position of the mirror relative to the null corrector must be known to an accuracy better than the tolerances on those parameters (for LBT, 1 mm for both radius of curvature and centration). The certifying hologram is a valuable tool for this measurement because it provides an optical reference to the test wavefront (the diffraction pattern on the hologram) as well as a mechanical reference for the distance measurement (the glass substrate on which the pattern is written).

The axial displacement between the mirror and the hologram can be measured with a steel tape or a laser tracker. The laser tracker has become an indispensable tool for measurement of large optics. It combines a distance-measuring interferometer with angular encoders and a steering mechanism to follow a moving retroreflector and measure its position in 3 dimensions. It includes an absolute distance measurement accurate to tens of microns over the distances involved in measuring large optics. Radial displacements (along the interferometer's line of sight) are measured to submicron accuracy and angles to better than 1 arcsec. Including all sources of uncertainty, the radius of curvature can be measured to an accuracy of a few tenths of a mm.

When the mirror is aligned to minimize coma in the interferometric test, its optical axis matches that of the test wavefront. This will generally be different from the mirror's mechanical axis. The relation between the two axes can be found by rotating the mirror on its mechanical

axis and measuring the change in coma. The displacement of the mechanical axis can be controlled to a few tenths of a mm with dial indicators or a laser tracker, and the coma can be measured to an accuracy better than 100 nm rms, which for LBT corresponds to 0.16 mm of displacement.

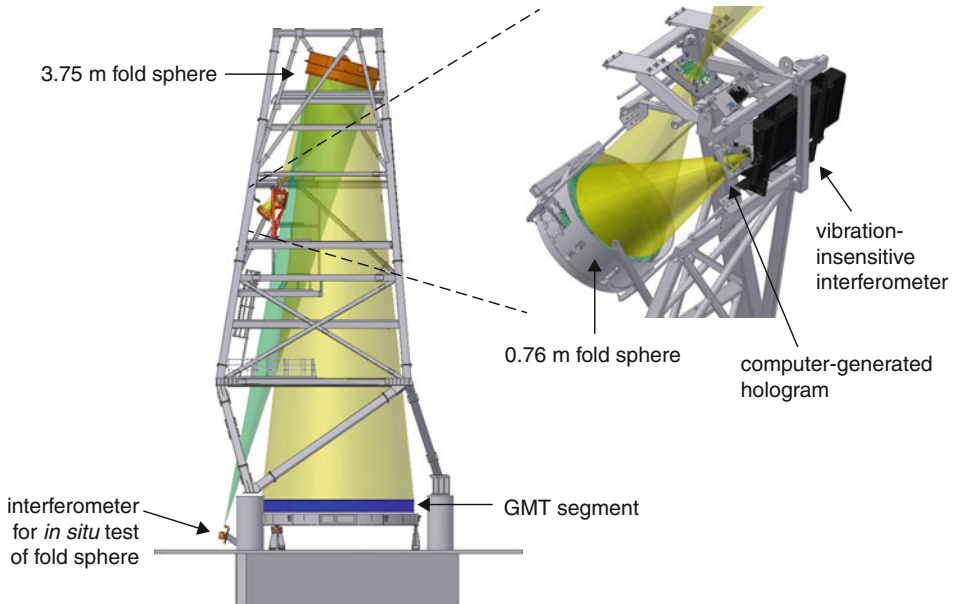
The conic constant k represents spherical aberration in the surface. The amount of spherical aberration in the template wavefront is determined by the null corrector. All parameters of the null corrector must be controlled to an accuracy consistent with the tolerance on conic constant or spherical aberration (for LBT, 0.001 in k or 46 nm rms spherical aberration). Since the certifying hologram validates the accuracy of the null corrector, it must be accurate to the same tolerance in spherical aberration.

Measuring the mirror surface to an accuracy of nanometers on small scales and around 100 nm on large scales requires interferometry at visible wavelengths, which can be done only for a polished surface. Before the surface is fully polished, it must be measured to an accuracy one to two orders of magnitude looser in order to limit the figure change that will have to be achieved by polishing. Interferometry at infrared wavelengths, typically with a 10.6 μm CO₂ laser, works with a ground surface and has adequate accuracy and resolution. It generally requires a separate null corrector using IR-transmissive elements. Another option is to scan the surface with a laser tracker (Zobrist 2009). The laser tracker is most accurate in its radial dimension, so the best surface measurement is obtained when the tracker is located near the mirror's center of curvature, and its line of sight is roughly normal to the surface. The measurement is sensitive to drift in the position of the mirror and the laser tracker during the scan. Such motion can be measured and compensated by continuously monitoring fixed retroreflectors on the mirror. The laser tracker method is attractive in cases where an IR null corrector is difficult or impractical, as is the case for the GMT segments (Zobrist et al. 2010).

The mirror's shape depends on its support forces and temperature distribution, so both support and temperature must be controlled for the lab measurements. For honeycomb mirrors, the support in the lab is nominally identical to that in the telescope at zenith pointing but is achieved with a passive hydraulic system instead of the telescope's active pneumatic system. Both systems have load cells to measure the force at each of the roughly 160 supports. The stiff mirrors can tolerate maximum support errors of about 20 N without significant impact on the figure apart from low-order bending modes that will be set by the active optics system at the telescope.

The borosilicate honeycomb mirrors are relatively sensitive to temperature gradients. Thermal control of the mirror is available in the lab as well as the telescope, but a simple system – drawing ambient air across the mirror's rear surface – is adequate because of the stable conditions in the lab. Roughly 100 thermocouples are attached to the internal surfaces of the honeycomb and monitored. Analysis shows that, for most temperature distributions, departures up to 0.05 K from an isothermal mirror have no significant effect on mirror figure. Somewhat tighter tolerances apply for axisymmetric temperature gradients. This is especially important for the GMT segments because they have to match in radius of curvature. An axial gradient of 0.05 K p-v changes the radius by about 0.4 mm, roughly the level of uncertainty that can be tolerated. The gradients can be kept below this level in the lab. The thermocouple system allows measurement of axial gradients down to a level of about 0.01 K, and a correction based on analysis can be applied to the measured figure.

The off-axis GMT segments present challenges beyond those of large symmetric telescope mirrors. Each segment is ten times more aspheric than an LBT primary mirror, and the aspheric departure is not axisymmetric. The null corrector for the interferometric test



■ Fig. 4-11

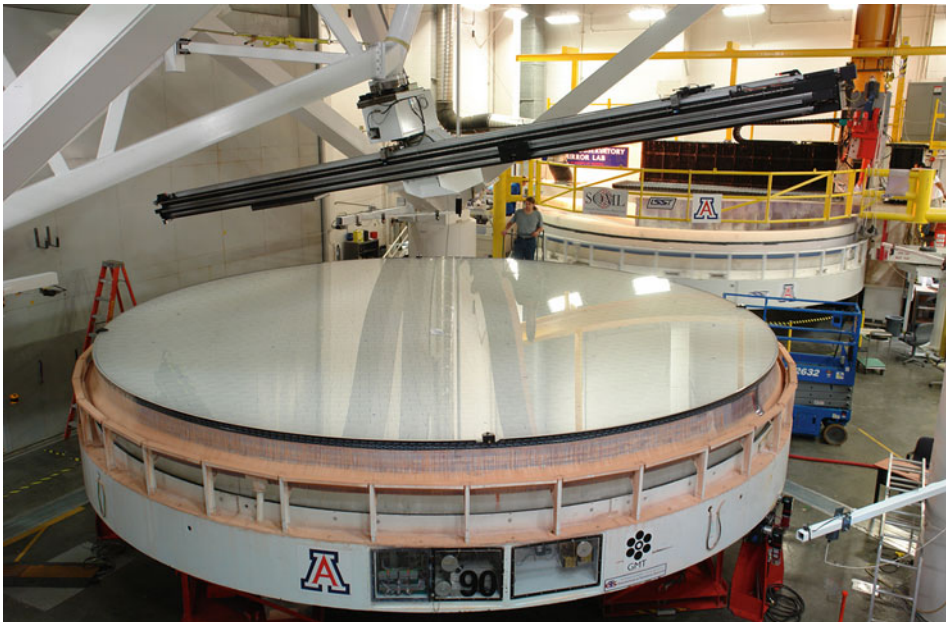
Model of the principal optical test for the GMT off-axis segments, in the 28 m test tower. At the right is a blow-up of the interferometer and first two elements of the null corrector. The reference hologram is inserted for alignment of the test system but removed for the measurement of the GMT segment. Gold light cones represent the measurement of the GMT segment, while the aqua cone in the full model at left represents a simultaneous measurement of the 3.75 m fold sphere

introduces the asphericity in the test wavefront using a large non-axisymmetric optical system shown in [Fig. 4-11](#) (Burge et al. 2008). It consists of a 3.75-m spherical mirror, a 75-cm spherical mirror, and a computer-generated hologram. Oblique reflections off the two mirrors do most of the shaping of the wavefront, and the hologram cleans up the remaining aberrations. In addition to introducing most of the aspheric departure, the large mirror folds and shortens the test beam so it can fit in the 28-m test tower.

The most challenging aspect of the GMT null corrector is alignment. Small misalignments cause large amounts of focus, astigmatism, and coma, and the lack of symmetry complicates the alignment process. The small package containing the interferometer, hologram, and 75-cm mirror requires an alignment accuracy of about $10\ \mu\text{m}$, and the larger dimensions between that package, the 3.75-m mirror, and the GMT segment have to be controlled to about $100\ \mu\text{m}$. This accuracy is obtained through use of computer-generated holograms and laser trackers. Like the certifying holograms described earlier, a reference hologram can be aligned optically to the wavefront and provides both optical and mechanical references so other components can be aligned to the wavefront. A laser tracker is used to measure the position and orientation of the reference hologram as well as the components of the test system (West et al. 2010).

For symmetric mirrors, it is possible to validate the null corrector with a certifying hologram. The GMT test wavefront is over 3 m in diameter by the time it leaves the null corrector, much too large to validate with a hologram. But the same goal can be achieved with an

independent measurement of the segment, in particular a measurement of the low-order aberrations that are sensitive to misalignment of the null corrector. For this purpose Burge and Su (Su et al. 2008, 2009) developed a scanning pentaprism system that scans the surface with a narrow collimated beam, which is focused on a detector in the mirror's focal plane. The displacement of the focused spot is proportional to the slope error on the surface. The scanning pentaprism test is accurate to about 0.1 arcsec rms surface slope. Two features of the test account for this remarkable accuracy. First, the pentaprism deflects an incoming beam in a fixed direction (in one dimension) independent of the orientation of the pentaprism. This keeps the beam parallel to the optical axis as it scans across the surface. Second, the system includes a nonmoving pentaprism beam splitter that produces a second spot in the focal plane. Displacement common to both spots indicates misalignment or instability of components including the light source, the GMT segment, and the focal plane detector, while differential motion between two spots is proportional to slope error of the mirror surface. A set of four scans across different diameters of the segment determines the first eight low-order aberrations to an accuracy similar to that of the interferometric test and well within the correction range of active optics at the telescope. Agreement between the interferometric test and the pentaprism test provides high confidence that all segments will perform at the specified accuracy. ▶ [Figure 4-12](#) shows the scanning pentaprism system for the GMT segments.



■ Fig. 4-12

GMT1 segment with the 8.4-m pentaprism rail under the optical test tower. A pentaprism on the rail is used to scan a collimated beam across the segment. The rail lies in a plane perpendicular to the optical axis of the parent mirror, making the collimated beam parallel to the axis. Not shown is the pentaprism detector, which is at the parent's focus 18.5 m above the segment and 4.5 m to the right of the segment center. (Steward Observatory photo by R. Bertram)

4 Experience in Operating Astronomical Telescopes

4.1 SDSS 2.5 Meter

The telescope built by the Sloan Digital Sky Survey (SDSS) on Apache Point, New Mexico, is dedicated to an imaging and spectroscopic survey of the sky (York et al. 2000). The 2.5-m $f/2.25$ borosilicate primary mirror was cast by Hextek and polished by the Optical Sciences Center at the University of Arizona. The $f/5$ secondary mirror with a 3° field of view is a gas fusion blank also made by Hextek and polished by SOML. The telescope is instrumented for wide-field CCD photometry and plug-plate fiber spectroscopy. The survey described by Abazajian et al. (2003) and references therein has been tremendously successful. Science results include discovery of a number of dwarf spheroidal companions to the Milky Way (Belokurov et al. 2007) and young galaxies at redshifts above 6 (Fan et al. 2004).

4.2 Lennon 1.8 Meter

The Alice P. Lennon telescope was built in 1990 on Mt. Graham, Arizona, by the Vatican Observatory and the University of Arizona. Its 1.8-m $f/1.0$ primary was the first spincast blank produced by the Mirror Lab (Goble et al. 1985; Martin et al. 1992). The short focal length optics and the altitude-azimuth mounting lead to an extremely compact mechanical structure and dome and the unofficial moniker of Vatican Advanced Technology Telescope (VATT). A Gregorian $f/9$ secondary feeds CCD imagers, a polarimeter, and a low-resolution spectrograph. Science observations include a microlensing survey of stars in M31 reported by Uglesich et al. (2004). VATT was the first optical telescope installed on Mt. Graham.

4.3 APO 3.5 Meter

The 3.5-m telescope built by the Astrophysical Research Consortium at Apache Point Observatory (APO) in New Mexico uses the first 3.5-m $f/1.75$ honeycomb blank produced by the Mirror Lab (Goble et al. 1989). The secondary and tertiary mirrors are gas fusion blanks from Hextek. The telescope uses a variety of spectroscopic and imaging instruments at an $f/10$ Cassegrain focus and is automated for remote observations. Many follow-up observations have been made of objects discovered with SDSS. The telescope mount and its performance have been described by Mannery et al. (1989).

4.4 WIYN 3.5 Meter

The WIYN 3.5-m telescope was constructed on the southwest ridge of Kitt Peak, Arizona, by a consortium including Wisconsin, Indiana, Yale, and NOAO (WIYN). It uses the second 3.5-m $f/1.75$ borosilicate honeycomb mirror produced by the Mirror Lab (Anderson et al. 1994). It was initially built as a spectroscopic telescope to provide a wide field of view for the workhorse HYDRA fiber spectrograph (Barden and Armandroff 1995). Typical multifiber science results with HYDRA involve measuring redshifts of stars or galaxies in clusters (Miller et al. 2004). The compact and light weight 3.5-m telescope is a notable contrast to the massive Mayall 4-m

telescope built on Kitt Peak 20 years earlier. The careful thermal design of the telescope and enclosure quickly demonstrated that very good imaging was routinely possible on Kitt Peak. Active optics using the primary mirror support actuators were included in the telescope design from the beginning (Roddier et al. 1995).

4.5 SOR 3.5 Meter

The Air Force Phillips Laboratory installed the 3.5-m telescope at its Starfire Optical Range (SOR) on Kirtland Air Force Base near Albuquerque, New Mexico, in 1993. The $f/1.5$ honeycomb primary mirror was cast and polished by SOML. The polishing and the active optics system of the primary mirror have been described by Martin et al. (1994). The high-speed mount was built by Contraves USA in Pittsburgh, and the retractable enclosure was built by Coast Steel in Vancouver. The telescope facility has been described by Fugate (2003). The telescope is equipped with a high-performance adaptive optics system and laser guide star. Adaptive optics results with a 50-W sodium laser guide star have been presented by Denman et al. (2006). The telescope is typically used for imaging man-made earth-orbiting satellites and for the development of imaging technologies.

4.6 MMT 6.5 Meter


The original MMT with six 1.8-m telescopes on a common mount was highly successful in terms of providing a 4.5 m effective collecting area with the resolution of a 6.9-m telescope and as a platform for instrument development. It was the first of a generation of new large telescopes on computer-controlled elevation-over-azimuth mounts. Astronomers were eager to replace the original six primary mirrors with a 6.5 m honeycomb mirror when the opportunity became available because the converted MMT would have 2.4 times the light-gathering power and a very similar resolution. The first 6.5 m borosilicate honeycomb mirror was cast by SOML in 1992 (Hill and Angel 1992). The MMT Conversion Project has been described by West et al. (1997) with the new telescope going on-line in 2000 (Blanco et al. 2004). This was the first of the large honeycomb mirrors to go into operation on sky. The $f/1.25$ primary mirror meant that a new elevation structure could fit on the existing azimuth yoke in the existing enclosure with only minor modifications. Martin et al. (1998a) describe the figuring of the fast paraboloidal primary mirror. The active supports and force optimization for the primary mirror have been described also by Martin et al. (1998b). The new MMT has three Cassegrain focal ratios to feed a variety of instrumentation. An $f/9$ secondary (Hextex honeycomb) feeds several legacy instruments developed for the old MMT. A 1.7-m $f/5$ secondary (machined Zerodur honeycomb) combined with a refractive corrector provides a 1° field-of-view for imaging and multifiber spectroscopy (Hastie and McLeod 2008). An $f/15$ adaptive secondary (Brusa et al. 2004) feeds a new generation of infrared instrumentation compatible with Magellan and LBT. This adaptive secondary with 336 actuators has been operating at the MMT for several years and is a precursor to the units now operating on LBT. Typical science observations are similar to those of Liu et al. (2007) who used adaptive optics and nulling interferometry to resolve disks around nearby Herbig Ae stars. A Rayleigh laser guide star system has been developed by Milton et al. (2008). The MMT is also the first telescope to do in situ aluminizing of the primary mirror.

The logic was that it was easier to bring the vacuum belljar to the mirror than it was to move the mirror to the belljar. The partners in the MMT are Arizona and Harvard/Smithsonian.

4.7 Magellan 6.5 Meter

The twin 6.5-m Magellan telescopes (Shectman and Johns 2003) named for Walter Baade and Landon Clay are located at Las Campanas Observatory in Chile. Each separate telescope has a 6.5-m honeycomb primary mirror with support technology similar to MMT and LBT. The figuring of the $f/1.25$ primary mirror and its support system with 104 active supports has been described by Martin et al. (2000). An excellent site with good thermal control of the telescope structures and mirrors leads to excellent image quality with images as sharp as 0.2 arcsec (Osip et al. 2008). The active optics system for Magellan with Shack-Hartmann wavefront sensors has been described by Schechter et al. (2003). A classical Gregorian optical design provides a wide field of view at $f/11$. The IMACS instrument (Dressler et al. 2006) is able to do imaging and multislit spectroscopy over a 27-arcminute field of view. Herbert-Fort et al. (2010) used IMACS to study the kinematics of nearby galaxy NGC 628. Some of the $f/5$ and $f/15$ instrumentation at Magellan is shared with the MMT. An $f/15$ adaptive secondary mirror for Magellan is under development. The Magellan primary mirrors are aluminized in a central facility between the two telescopes. The partners in the Magellan Telescope are the Carnegie Institution of Washington, the University of Arizona, Harvard University, Massachusetts Institute of Technology, and the University of Michigan.

4.8 LBT 2 × 8.4 Meter

The Large Binocular Telescope (LBT) uses two 8.4-m-diameter honeycomb primary mirrors mounted side by side on the same mount to produce a collecting area (110 m^2) equivalent to an 11.8-m circular aperture. The two Gregorian telescope sides point at the same field. A unique feature of LBT is that the light from the two primary mirrors can be combined optically in the center of the telescope to produce phased array imaging of an extended field. In practice, this extended phased field can be of order 1 arcmin in diameter. Active and adaptive optics have been designed into the telescope from the beginning to augment the telescope performance from visible to mid-infrared wavelengths. The main wavefront correctors are the two $f/15$ Gregorian adaptive secondary mirrors. The interferometric focus combining the light from the two 8.4-m primaries will reimage the two folded Gregorian focal planes in a central location. Several of the instruments will implement an additional wavefront corrector at a higher conjugate after the initial Gregorian focus. This cophased imaging gives the telescope the diffraction-limited resolution of a 22.65-m telescope in one spatial direction. Images at the combined focus have a resolution of 5 milliarcsec in visible light and 20 milliarcsec in the near infrared. The binocular configuration of the telescope optics leads to a compact and stiff mechanical structure. The fast primary mirrors help minimize the size of the corotating enclosure.  Figure 4-13 shows the telescope enclosure open in twilight for observing. The two $f/1.14$ primary mirrors of LBT are mounted on an altitude-over-azimuth mount with hydrostatic bearings operating at 120 bar. There is a quite direct load path from the primary mirror cells to the elevation bearings to the azimuth structure to the azimuth bearings to the telescope support pier. The azimuth and elevation drives each have four brushed DC servomotors directly



■ Fig. 4-13

Photo of the Large Binocular Telescope with two 8.4-m- diameter primary mirrors open and ready for observing at sunset with the two prime focus cameras (deployed) and two Gregorian secondary mirrors (retracted). The telescope enclosure with two 10-m wide shutter openings co-rotates with the azimuth axis of the telescope at speeds up to $1.5^\circ/\text{sec}$. The forest around the telescope on Mt. Graham is composed of spruce, fir, and aspen trees. The trees on the hill in the background burned during a forest fire in July 2004

driving pinions (60:1) on common shafts against large diameter segmented gear sectors. The drives in elevation and azimuth are applied at a 7-m radius to give high stiffness as described by Davison (1990). This stiff mechanical design combined with the light weight optics allows the LBT to achieve a minimum eigenfrequency of nearly 8 Hz. The telescope is located at Mount Graham International Observatory in the Pinaleño Mountains of southeast Arizona at an elevation of 3,192 m. The development of the LBT has been previously described by Hill et al. (2008b), and by Hill (2010) and by Hill et al. (2010).

The two paraboloidal primary mirrors can also be used independently (without optical combination) to obtain seeing limited images over a wide field of view. The LBT offers instruments mounted at a variety of focal stations. Instruments with wide fields of view, including imagers and low-moderate resolution spectrographs, are mounted in pairs at the separate foci of the two telescopes. These instruments are typically background limited and gain little benefit from optical combination of the light from the two sides of the telescope. Interchange between

the various optical configurations during the night is accomplished with swing arms that hold the secondary mirrors, the tertiary mirrors, and the prime focus correctors. Monocular prime focus science imaging started in fall 2006, with regularly scheduled science observations using both primary mirrors in parallel at prime focus starting in January 2007. The prime focus cameras have been described by Ragazzoni et al. (2006). An early science result was the deep imaging of the Hercules dwarf galaxy by Coleman et al. (2007) which showed the galaxy to be elongated. Much of the deep imaging time with the prime focus cameras has been devoted to adding U-band and z' -band imaging to deep survey fields thereby taking advantage of the large collecting area of LBT. As an example, see the deep near-UV galaxy counts reported by Grazian et al. (2009). Commissioning of the multiple $f/15$ Gregorian focal stations began in April 2008 and continues through the present. The initial observations used a conventional secondary mirror of Hextek design mounted on the left side of the telescope along with a Hextek tertiary flat mirror. The first science observations at the bent Gregorian focus began in December 2009. Bian et al. (2010) used LUCI1 to obtain images and infrared spectra of a redshift two galaxy lensed by a foreground cluster. A near-infrared image of the planetary nebula M57 obtained with LUCI1 is shown in [▶ Fig. 4-14](#).

Commissioning of the adaptive optics system on sky began in May 2010 and is discussed below in [▶ Sect. 4.8.6](#). Instruments that work at the diffraction-limit benefit significantly from the increased resolution obtained by combining light from the two sides of the telescope (Hill 1994; Angel et al. 1998). The first cophased binocular experimental observations were in October 2010 (see the discussion of phasing and interferometric instruments below under [▶ Sect. 4.8.7](#).)

4.8.1 Support and Positioning of Primary Mirrors

Each primary mirror is supported in the LBT on a set of 160 loadspreaders which are glued to the flat backplate of the mirror substrate. A pneumatic force actuator connects to each of the loadspreaders in order to provide a force to float the mirror against the changing gravity and wind loads. Both axial and lateral forces are applied only to the backplate of the glass honeycomb. One hundred and four of the actuators have a second cylinder angled at 45° from the axial cylinder to provide lateral forces in vector combination. Four actuators have the second cylinder rotated into the cross-lateral plane (parallel to the elevation axis) to stabilize side-to-side motion. The force actuators are controlled by individual servos which allow active correction forces to be applied to adjust the mirror figure in addition to pre-calculated forces to compensate gravity. These actuators apply forces that are independent of the mirror position (relative to the mirror cell) over a range of 8 mm in all directions. A set of static supports holds the mirror in the cell in case either power or air pressure is lost.

Six hardpoints arranged as a Stewart platform define the kinematic position of each primary mirror in its cell without carrying a significant portion of the load. These hardpoints are adjustable in length to provide the collimation and focussing (6 degrees of freedom in position) of the mirror in the telescope. Each of the hardpoints includes a breakaway mechanism so that they cannot apply large local forces to the back of the mirror. The forces on these stiff hardpoints are measured with a set of loadcells and used to adjust the overall set of commands to the force actuators so that the load on each hardpoint is minimized (servoed to a small fixed offset). The six hardpoint forces are resolved into three net forces and three net moments that can be compensated with corresponding sets of axial and lateral support forces (applied by the pneumatic



■ Fig. 4-14

This infrared image of the Ring Nebula (Messier 57) shows molecular hydrogen emission in red as a tracer of the shock waves from distinct mass ejection episodes, and Brackett-gamma emission in blue from Hydrogen ionized by the central star. Continuum emission at $2.2\ \mu\text{m}$ displayed as green contributes to the traditional ring that is also observed in visible light (LBTO image courtesy of D. Thompson)

actuators) with minimal effect on the mirror figure. Each force actuator receives a feed-forward force command based on the telescope elevation to float the mirror against gravity. The outer servo loops provide additional force commands to react to wind, telescope acceleration, and residual gravity loads as sensed by the hardpoints. The mirror support concept and the actuator design were described by Gray et al. (1994). Additional performance details of the LBT mirror support system are discussed by Ashby et al. (2008).

4.8.2 Active Optics and Wavefront Sensing

Active optics is principally used for correcting the alignment and low-order figure of the primary mirrors plus the telescope focus and collimation by making slow adjustments to the mirror position and support forces. These adjustments allow the compensation of residuals from the gravity supports, the wind load, and thermal effects in the mirror and telescope structures. Corrections are typically made on 1-min timescales which are slower than atmospheric

turbulence but much faster than thermal and gravitational changes in the telescope. Even thick and relatively stiff mirrors are able to benefit from active correction although the total range of adjustment is smaller (Parodi et al. 1992). Minimal additional complexity is added to the mirror support system other than to require that the forces on all actuators be adjustable under computer control.

The strategy for the support of the plano-concave honeycomb mirrors is to compute the required support forces to float the mirror against gravity using finite element techniques (Parodi et al. 1997). The mirrors are polished on a hydraulic support system that applies the calculated forces at the zenith-pointing position. This means that first-order deviations from the ideal support forces are polished out from the zenith figure and only appear in reverse at horizon. After polishing, the mirror is mounted in the telescope support cell while still located under the test tower in the optical lab. The precision optical test is then used to confirm that the telescope supports match the polishing supports and to tune out small residual bending errors. As part of the optimization process, a number of bending modes were measured and compared with the calculated modes (Martin et al. 2004). The forces supplied by the active pneumatic actuators in the telescope support cells are calibrated to a part in 1,000. This precision calibration is not strictly necessary when the telescope uses active optics, but it allows failed actuators to be replaced with spare units without significantly changing the figure of the mirror. Using the calibrated force actuators, the figure of the LBT primary mirrors is stable and matches that measured in the Mirror Lab 5 years earlier to the level of 1 arcsec images without any active optics correction. Residual errors are dominated by a few hundred nm of astigmatism from stray forces. The mirrors are stiff enough that only the axial forces (along the optical axis) need to have active adjustment. The lateral forces remain at their pre-computed values (varying with elevation) while operating in the telescope.

Three kinds of wavefront sensing are used on LBT (through 2010). At prime focus, extra-focal pupils provide focus, collimation, and low-order active figure correction to the primary mirrors. Typically, this CCD sensor slightly displaced from the focal plane takes 16-s exposures and is used every 20–40 min to update the collimation. Lookup tables correct the collimation and focus of each primary mirror between the active corrections. This extra-focal pupil system is described by Hill et al. (2008b).

For seeing limited observations at the Gregorian foci, an off-axis guide probe uses a Shack-Hartmann wavefront sensor with 13×13 lenslets across the pupil. This sensor typically takes 30-s integrations (to average out the atmosphere) at 1-min intervals for making active optics corrections to the telescope collimation and the primary mirror figure. The off-axis guiding and wavefront sensing units have been described by Storm et al. (2004). Active corrections are typically made once per minute to account for gravity-induced flexure and thermal disturbances. With predetermined lookup tables to follow the changes, the active correction interval can be reduced to a few times per hour for seeing-limited observations. On a stable night, the active optics residuals are less than 100 nm rms wavefront for Zernike terms Z4–Z11, Z22 (focus, astigmatism, coma, trefoil, plus third- and fifth-order spherical) to give telescope images that are dominated by seeing at 0.3 arcsec FWHM.

The adaptive optics system uses an on-axis pyramid sensor to measure the wavefront at up to 1-kHz rate. The red light from a natural guide star is reflected off a tilted dichroic window in front of the infrared science instrument. To reduce the bending forces on the adaptive secondary shell, certain low-order wavefront errors are off-loaded to the positions of the mirrors and the shape of the primary on a timescale of 10–20 s.

4.8.3 Ventilation and Thermal Management

The honeycomb primary mirrors on LBT are actively air-conditioned to follow the changing ambient temperature in the nighttime environment of the telescope. This conditioning allows the honeycomb structure to follow the mountain ambient temperature with a thermal time constant of 45 min. It minimizes distortion of the mirror surface due to temperature gradients and minimizes mirror seeing due to convection when the mirror surface deviates from ambient temperature. Cheng and Angel (1986) found that 8 l/sec of air circulated through each honeycomb cell kept the mirrors in good thermal equilibrium. Air entrainment devices (a.k.a. jet ejectors) are used in LBT to pressurize air inside the mirror support cell and to circulate air through the glass honeycomb. In practice, this ventilation system works well, and the mirrors are normally within a fraction of a degree of ambient in the observing chamber. The internal temperature gradients across the glass are approaching 0.1°C when the temperature is changing slowly, although gradients at this level are challenging to measure. Thermal errors in the primary mirror shapes are within the force correction range of the active optics system.


The LBT also does some thermal management of the telescope structure and the observing chamber to control the dome seeing. Most of these strategies for seeing control were summarized by Bely and LeLievre (1987). All instrumentation on the telescope is liquid-cooled to keep it near ambient temperature (but still above the dew point) and to avoid warm surfaces in contact with the air within the telescope. Structural members of the telescope with thin cross sections are wrapped with low emissivity aluminum foil tape to minimize the radiative cooling to the cold night sky. Thicker sections of steel (50–100 mm) such as the azimuth frame and the elevation sectors will be covered by a sheet metal skin to control any warm air created in that area (this “stealth” air system has not yet been completed on the telescope). Ambient air will be drawn in behind this metal skin with fans and the warm air exhausted downwind (Salinari and Hill 1994). The heavy steel columns of the rotating enclosure are fabricated as trusses to control the thermal time constant of the enclosure interior (Teran et al. 2002). Unused daytime chiller capacity can be used to cool the observing chamber. The air circulation system on the roof for melting accumulated snow also has the ability to remove daytime heat from the roof skin. The roof surface is coated with aluminum tape to minimize radiative cooling during the night. Large vent doors at the sides and rear of the observing chamber serve to flush ambient air through the observing chamber and telescope structure on nights when the wind is low.

4.8.4 Reflective Coating

Like MMT, the LBT was designed to aluminize the two 8.4-m primary mirrors in situ on the telescope structure. After washing and chemically stripping off the old aluminum, the telescope is moved to the horizon-pointing position and locked in place. The 25-ton aluminizing belljar is lifted up with the overhead crane and sealed against the corresponding vacuum flange on the primary mirror cell. The technical details and results of the aluminizing system are described by Atwood et al. (2006). The rough vacuum to 20 mTorr is made with a roughing pump and a Roots blower. Inside the belljar, a group of charcoal cryopanel cooled by liquid nitrogen provides the final pumping down to the coating vacuum of $6\ \mu\text{Torr}$. Aluminum is evaporated by a set of 28 BN crucibles heated with tungsten heaters. The LBT system is unique in that each of these 28 source modules contains an integrated transformer mounted with the crucible heater.

The control system supplies 280 VAC at 20 kHz which is then stepped down in voltage and up in current by a factor of 26 at the location of the crucibles. Each of the first seven primary mirror coatings has been successful as evaluated by reflectivity and coating thickness measured on witness slides. Selecting in situ aluminizing was a good choice operationally, and the innovative design of the source modules with transformers has been a technical and economic success. The only significant penalty to the telescope design was the need to make the mirror cell a vacuum vessel. This makes the cell weldment somewhat heavy (44 tons), restricts access to the mirror supports in some areas, and requires mechanisms in the cells to use vacuum-compatible lubrication.

4.8.5 Handling and Transportation

The transportation of the 8.4-m mirrors from the SOML polishing facility in Tucson to the mountaintop was a cooperative effort between the University of Arizona and Precision Heavy Haul of Phoenix, Arizona. Moving a single large piece of glass is clearly one of the challenges of this primary mirror design. Each 8.4-m mirror was transported in a specially designed box which provides multiple layers of spring isolation. The mirror box was transported horizontally down interstate highway 10 between Tucson and Safford, Arizona. The truck with the mirror box traveled at speeds up to 80 km/hr on the highway with a rolling roadblock escort to control traffic. After arriving at the observatory basecamp, the box was tilted by 60° and mounted on a Goldhofer trailer for the drive up the 48 km of winding mountain road as shown in  Fig. 4-15. The mirror transport took 3 days of travel to reach the mountaintop site from Tucson. Davison et al. (2004b) provide additional details on the engineering aspects of mirror transportation and mirror handling.

4.8.6 Adaptive Optics

LBT is one of the first telescopes to have adaptive optics to compensate atmospheric turbulence designed in from the start. The main deformable elements in the LBT adaptive optics system are the secondary mirrors. Each of two concave ellipsoidal secondary mirrors on LBT is a 911-mm diameter Zerodur shell which is 1.6 mm thick. On the back side of each shell are 672 small magnets which can be actuated by DSP-controlled voice coils. The magnets and voice coils provide forces to bend the shell relative to a rigid reference body at 1,000-Hz rates. Capacitive sensors provide a reference signal for the fast servos that stabilize the shell position. The aspheric Zerodur (glass ceramic) shells of the adaptive secondaries are manufactured by the Steward Observatory Mirror Lab, while the mechanics are manufactured by ADS International in Lecco, Italy, and the electronics are manufactured by Microgate in Bolzano, Italy. The secondary mirror is made the deformable element in order to reduce the telescope background at longer wavelengths by having fewer warm optics (Lloyd-Hart 2000). The Gregorian optical design of LBT places the secondary shell conjugate to turbulent layers 100 m above the telescope. The ellipsoidal mirror also provides access to two real conjugate foci, so it is straightforward to test the optical surface of the secondary mirrors without light from stars. The first adaptive secondary unit was mounted on the right side of the telescope during March 2010, and with a static shape (nonadaptive) was able to deliver 0.4 arcsec FWHM images in red light (2-s exposures in r'). The second adaptive unit for the other side of the telescope is at Arcetri

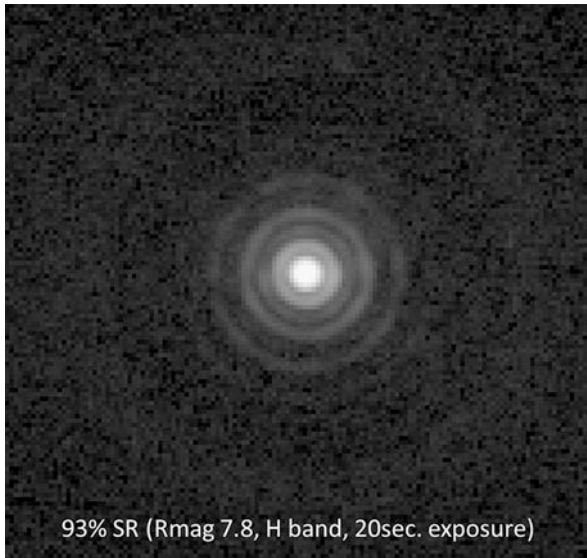


■ Fig. 4-15

This photo shows the first LBT 8.4-m primary mirror on the way up the mountain road on Mt. Graham in 2004. The mirror transport box is mounted on a Goldhofer trailer with a tractor pulling and a loader pushing. An operator on the Goldhofer controls the lateral tilt of the load as the road slope changes. The trip up the 48-km mountain road took 2.5 days while stopping at night to rest

Observatory for closed-loop system testing before moving to the telescope in 2011. Each adaptive system is capable of providing a Strehl ratio above 90% at a wavelength of $1.6\ \mu\text{m}$. The LBT adaptive optics system uses an on-axis pyramid sensor to measure the atmospheric wavefront at up to 1-kHz rate using a natural star. The wavefront sensor units have been designed and constructed at the Arcetri Observatory in Firenze, Italy. A Rayleigh laser system to allow correction of ground-layer turbulence over an extended field (4 arcmin) is under development (Rabien et al. 2010).

The adaptive optics system on the right side of LBT achieved better than 60% Strehl ratio on sky in H-band (400 modes, 1,000 frames/sec) on the first night of adaptive optics commissioning in May 2010. The initial performance was limited by residual vibrations in the telescope and some calibration issues. (See Esposito et al. (2010) for additional details of the laboratory acceptance tests and on-sky commissioning.) Recent data shows an H-band Strehl of 93% when correcting on a bright star with 500 modes as shown in ► Fig. 4-16. This excellent performance is due to a combination of a continuous deformable mirror with 672 actuators, an efficient pyramid wavefront sensor, a primary mirror which is smooth at mid-scales, and careful calibration of the system in the laboratory. The LBT mount tracks open loop to a precision of 10 milliarcsec



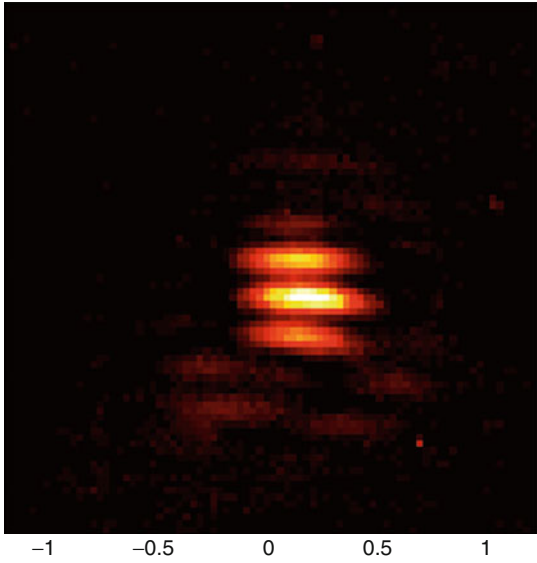
■ Fig. 4-16

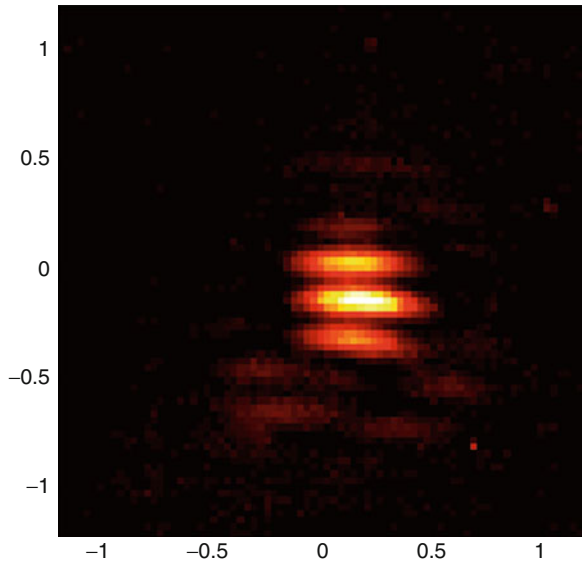
This image from the infrared test camera was taken during commissioning observations with the adaptive secondary on LBT in November 2010. The adaptive optics loop has been closed on a bright star with 500-mode correction and a 1,000-Hz frame rate. This 20-s H-band ($1.6\ \mu\text{m}$) exposure has a Strehl ratio of 93% and a FWHM in the diffraction-limited core of 40 milliarcsec. The halo outside of the diffraction rings is from wavefront errors at higher spatial frequency than the correction of the adaptive secondary. The image is displayed with a logarithmic stretch to make the diffraction rings clearly visible (image courtesy of the LBT Adaptive Optics team at Osservatorio Astrofisico di Arcetri)


rms to provide a stable base for adaptive optics. The adaptive system provides useful correction at lower Strehl ratio for stars as faint as magnitude 17. Additional information on adaptive optics in general can be found in [Sect. 4.8.6](#).

4.8.7 Phased Array Imaging

Two instruments to combine the light optically from the two sides of LBT are under development. These instruments are often called “interferometric” because the light from the two sides of the binocular telescope is combined together at a common focal plane in phase. Fizeau-mode beam combination combines the light in the focal plane while preserving the relative geometry of entrance and exit pupils in a homothetic manner in order to have an extended phased field. From the telescope point of view, the two telescopes are combined as a phased array which is a more optically precise description. The diffraction-limited PSF in the combined focal plane is the Airy pattern of a single 8.25-m aperture modulated by Young’s fringes

whose spacing depends on the center-center separation of the apertures (Sabatke et al. 2006). The LINC-NIRVANA instrument (Herbst et al. 2010) will work at 1–2.5 μm in the near infrared to do phased array imaging in Fizeau mode with multiconjugate adaptive optics. The 22.65-m baseline, B , of LBT will provide diffraction-limited resolution (λ/B) of 20 milliarcsec at 2 μm in one spatial direction. Several of these observations at different parallactic angles can be combined numerically to produce the Airy PSF of a circular aperture 22.65-m telescope (Carillet et al. 2002). The LBTI instrument (Hinz et al. 2008) will work at longer wavelengths for both Fizeau imaging and Bracewell nulling interferometry to image planets around nearby stars. During its first night of testing on sky in October 2010, LBTI obtained Fizeau fringes in the diffraction pattern at wavelengths of 5, 10, and 12 μm . These fringes in the PSF at 12 μm can be seen in  Fig. 4-17. Nulling interferometry observations will be made in the future with LBTI when adaptive optics are correcting both sides of the telescope. The nulling interferometer overlaps the two 8.25-m pupils on a 50% transmissive optic, with a half-wave phase step between the two beams, in order to project a pattern of light/dark fringes on the sky (rather than the focal plane). When a bright star is placed in the central dark fringe (destructive interference), the neighboring bright fringe (constructive interference) will be used to search for exoplanets and exo-zodiacal dust. Hénault (2010) provides an overview of various types of imaging and nulling interferometers.



 Fig. 4-17

This image shows the diffraction pattern of the LBTI instrument on LBT which combines the two sides of the telescope in a common focal plane. This is a 0.1-s integration at a wavelength of 12 μm taken in November 2010 using the MIRAC camera. Active and adaptive optics were only operational on one side of the telescope at that time so, some residual coma and astigmatism can be noticed around the central core. The spacing of the fringes at 12 μm is 0.17 arcsec. The entire image is 2.4 arcsec square (Steward Observatory image courtesy of P. Hinz and W. Hoffmann)

4.8.8 Partners

The international partners in the Large Binocular Telescope Corporation include Arizona (25%), Germany (25%), Italy (25%), Ohio State (12.5%), and Research Corporation (12.5%). The Arizona portion of the project includes astronomers from the University of Arizona, Arizona State University and Northern Arizona University. The German portion is represented by the LBT Beteiligungsgesellschaft which is composed of Max-Planck-Institut für Astronomie in Heidelberg, Zentrum für Astronomie der Universität Heidelberg, Max-Planck-Institut für Radioastronomie in Bonn, Max-Planck-Institut für Extraterrestrische Physik in Munich, and Astrophysikalisches Institut Potsdam. National participation in Italy is organized by the Istituto Nazionale di Astrofisica (INAF). Partners at individual institutions include the Ohio State University in Columbus, Research Corporation in Tucson, the University of Notre Dame, the University of Minnesota, and the University of Virginia. These partners have joined to form the Large Binocular Telescope Corporation in order to build and operate this telescope. Astronomers and engineers at all of these institutions are involved in building instruments and auxiliary equipment for the telescope.

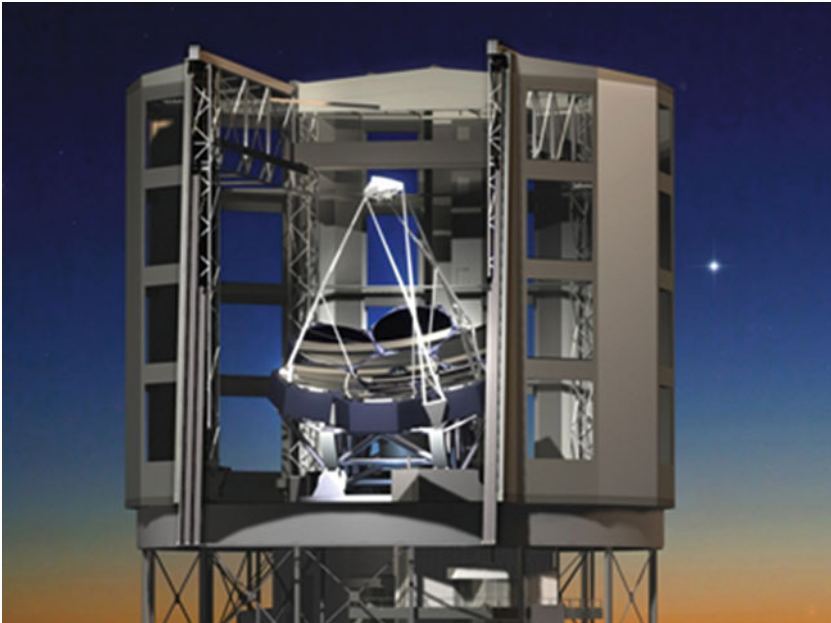
5 Future Telescopes Under Development

5.1 LSST

The Large Synoptic Survey Telescope (LSST), following a concept first proposed by Angel (Tyson et al. 2001), uses a 3-mirror design with an 8.4-m primary mirror to achieve a 3.5° field of view for surveying the sky quickly to a significant depth with a 6.5-m effective aperture. The primary mirror with a tertiary mirror polished into the same blank is being produced at SOML as described by Martin et al. (2008). The 3.4-m secondary mirror blank is being made by Corning from ULE. The survey has been described by Axelrod (2006), and the data management has been described by Kantor and Axelrod (2010). The design of the telescope system has been described by Krabbendam and Sweeney (2010). The telescope is a high performance design purpose-built for this particular survey (Davison et al. 2004). The detector package is a 3,200 Megapixel CCD array. The LSST is a US-led public data project with funding to be supplied by the National Science Foundation (NSF) and the Department of Energy (DOE). The telescope will be located in Chile at Cerro Pachón because of the frequently photometric skies and the available infrastructure.

5.2 GMT

The Giant Magellan Telescope (GMT) recently described by Sackett and Johns (2010) uses seven 8.4-m borosilicate honeycomb mirrors as circular segments of a 25-m $f/0.7$ primary. The first off-axis segment is under production at SOML as described by Martin et al. (2010). Many of the polishing and testing challenges are described in [▶ Sect. 3](#) above. The secondary mirror is divided into seven adaptive segments in the same geometry as the primary. The GMT relies on many of the technologies already demonstrated in the operational LBT including primary mirrors, adaptive secondary mirrors, and the telescope mount structure with large radius C-rings. The telescope will be located in Chile on Cerro Las Campanas as shown in [▶ Fig. 4-18](#).



■ Fig. 4-18

A model of the future Giant Magellan Telescope with seven 8.4-m segments making up a 25-m $f/0.7$ primary mirror. The GMT segments are the same size (8.4 m) as the LBT primary mirrors (compare this model with the photo of LBT in [Fig. 4-13](#))

Partners in developing the GMT include Carnegie Institution for Science, Harvard University, Smithsonian Astrophysical Observatory, Texas A&M University, Korea Astronomy and Space Science Institute, The University of Texas at Austin, The Australian National University, University of Arizona, Astronomy Australia Ltd., and The University of Chicago.

5.3 SASIR

The Mirror Lab cast a fifth 6.5-m $f/1.25$ honeycomb mirror in 2009 to become the primary mirror of a telescope to be located on San Pedro Martir, Baja California, Mexico. This telescope will be used in the Synoptic All-Sky Infrared Imaging (SASIR) Survey which is a collaboration between astronomers at Universidad Nacional Autónoma de México (UNAM), the University of California at Berkeley, Instituto Nacional de Astrofísica, Óptica and Electrónica (INAOE), and the University of Arizona.

Acknowledgements

The authors would like to thank Neville Woolf, Peter Strittmatter, Warren Davison, and J.T. Williams for their continuous wisdom, inspiration, and support over decades of mirror and

telescope development. Of course, the ideas discussed here could never turn into reality without the substantial engineering and organizational talents of the following persons at Steward Observatory: D. Anderson, R. Allen, J. Burge, B. Cuerden, S. DeRigne, L. Dettmann, K. Duffek, L. Goble, J. Hagen, S. Hinman, M. Hunten, D. Ketelsen, K. Kenagy, J. Kingsley, C. Kittrell, R. Lutz, S. Miller, D. Mitchell, R. Nagle, B. Olbert, B. Powell, B. Sisk, P. Schaller, B. Smith, M. Tuell, R. Warner, S. Warner, D. Watson, S. West, R. Young, C. Zhao, T. Zobrist, and numerous others.

The early funding of the Steward Observatory Mirror Lab came from the University of Arizona, from the National Science Foundation, and from the Air Force Weapons Laboratory. This present work has been supported by Steward Observatory and the Large Binocular Telescope Observatory at the University of Arizona.

References

- Abazajian, K., Adelman-McCarthy, J. K., Agueros, M. A., Allam, S. S., Anderson, S. F., Annis, J., et al. 2003, The first data release of the sloan digital sky survey. *Astron J*, 126, 2081–2086
- Allen, L. N., Keim, R. E., Lewis, T. S., & Ullom, J. R. 1992, Surface error correction of a Keck 10-m telescope primary mirror segment by ion figuring. *Proc SPIE*, 1531, 195–204
- Anderson, D. S., Martin, H. M., Burge, J. H., & West, S. C. 1994, Rapid fabrication strategies for primary and secondary mirrors at Steward Observatory Mirror Laboratory. *Proc SPIE*, 2199, 199–210
- Angel, J. R. P. 1984, Steps towards 8M honeycomb mirrors V. A method for polishing aspheres as fast as F/1. in *Very Large Telescopes, Their Instrumentation and Programs*, ed. M.-H. Ulrich, & K. Kjar (Garching: ESO), 11–21
- Angel, J. R. P., & Hill, J. M. 1981, Honeycomb mirrors of borosilicate glass. *Scientific Importance of High Angular Resolution at Infrared and Optical Wavelengths* (Garching: ESO), 61–65
- Angel, J. R. P., & Hill, J. M. 1982, Manufacture of large honeycomb mirrors. *Proc SPIE*, 332, 298–306
- Angel, J. R. P., & Hill, J. M. 1984, Steps toward 8m honeycomb mirror blanks III. 1.8m honeycomb sandwich blanks cast from borosilicate glass. *Proc SPIE*, 444, 194–199
- Angel, J. R. P., & Woolf, N. J. 1980, MT-2. *Optical and Infrared Telescopes for the 1990s* (Tucson: KPNO), 1062–1149
- Angel, J. R. P., & Woolf, N. J. 1984, Steps toward 8-meter honeycomb mirror blanks. I. Rationale and Approach. *Proceeding of the XI Texas Symposium on Relativistic Astrophysics* (Austin, TX: New York Academy of Sciences), 163–170
- Angel, J. R. P., Arganbright, D., Harmonson, L., Hill, J. M., & Woolf, N. 1981, Honeycomb mirrors of borosilicate glass: current results and plans for 7–8m diameter. *Instrumentation for Astronomy with Large Optical Telescopes*, ed. C. M. Humphries, 33–36
- Angel, J. R. P., Hill, J. M., Goble, L., & Woolf, N. J. 1983, Steps toward 8m honeycomb mirror blanks: IV. Structural design and the fabrication facility. *Proc SPIE*, 444, 194–199
- Angel, J. R. P., Davison, W. B., Hill, J. M., Mannery, E. J., & Martin, H. M. 1990, Progress toward making light weight 8m mirrors of short focal length. *Proc SPIE*, 1236, 636–640
- Angel, J. R. P., Hill, J. M., Strittmatter, P. A., & Weigelt, G. 1998, Interferometry with the large binocular telescope. *Proc SPIE*, 3350, 881–889
- Ashby, D. S., Kern, J., Hill, J. M., Davison, W. B., Cuerden, B., Brynneel, J. G., et al. 2008, The large binocular telescope primary mirror support control system description and current performance results. *Proc SPIE*, 7018, 70184C–12
- Atwood, B., Pappalardo, D. P., O'Brien, T. P., Hill, J. M., Mason, J. A., Belville, R., et al. 2006, The aluminumizing system for the 8.4-meter diameter LBT primary mirrors. *Proc SPIE*, 6273, 62730T–12
- Axelrod, T. S. 2006, The large synoptic survey telescope. *Astronomical Data Analysis Software and Systems XV*, Vol. 351 (San Francisco: Astronomical Society of the Pacific), 103–111
- Barden, S. C., & Armandroff, T. 1995, Performance of the WIYN fiber-fed MOS system: HYDRA. *Proc SPIE*, 2476, 56–67
- Beckers, J. M., & Williams, J. 1980, The MMT as it exists today. *Optical and Infrared Telescopes for the 1990s* (Tucson: KPNO), 108–126
- Beckers, J. M., Ulich, B. L., Shannon, R. R., Carleton, N. P., Geary, J. C., & Latham, D. W., et al. 1981, The multiple mirror telescope. *Telescopes for the 1980s* (Palo Alto, CA: Annual Reviews), 63–128

- Beckers, J. M., Ulich, B. L., & Williams, J. T. 1982, Performance of the multiple mirror telescope (MMT) I. – MMT the first of the advanced technology telescopes. *Proc SPIE*, 332, 2–8
- Belokurov, V., Zucker, D. B., Evans, N. W., Kleyna, J. T., Koposov, S., Hodgkin, S. T., et al. 2007, Cats and dogs, hair and a hero: a quintet of new milky way companions. *Astrophys J*, 654, 897–906
- Bely, P. Y., & LeLievre, G. 1987, Seeing control in domes and telescopes. in *Identification, Optimization, and Protection of Optical Telescope Sites*, ed. R. L. Millis, O. G. Franz, H. D. Ables, & C. C. Dahn (Flagstaff), 155–166
- Bian, F., Fan, X., Bechtold, J., McGreer, I. D., Just, D. W., Sand, D. J., et al. 2010, LBT/LUCIFER observations of the $z \sim 2$ lensed galaxy J0900+2234. *Astrophys J*, 725, 1877–1885
- Blanco, D., Alegria, M., Callahan, S., Clark, D., Commisso, B., Foltz, C. B., et al. 2004, The new MMT. *Proc SPIE*, 5489, 300–311
- Brusa, G., Miller, D. L., Kenworthy, M. A., Fisher, D. L., & Riccardi, A. 2004, MMT-AO: two years of operation with the first adaptive secondary. *Proc SPIE*, 5490, 23–33
- Burge, J. H. 1993a, Advanced techniques for measuring primary mirrors for astronomical telescopes. University of Arizona, Ph. D. dissertation, College of Optical Sciences
- Burge, J. H. 1993b, Certification of null correctors for primary mirrors. *Proc SPIE*, 1994, 248–259
- Burge, J. H., Davison, W., Martin, H. M., & Zhao, C. 2008, Development of surface metrology for the Giant Magellan Telescope primary mirror. *Proc SPIE*, 7018, 701814–12
- Carbillet, M., Correia, S., Boccacci, P., & Bertero, M. 2002, Restoration of interferometric images. II. The case study of the large binocular telescope. *Astron Astrophys*, 387, 744–757
- Cheng, A. Y., & Angel, J. R. P. 1986, Steps toward 8m honeycomb mirrors. VIII – Design and demonstration of a system of thermal control. *Proc SPIE*, 628, 536–544
- Coleman, M. G., de Jong, J. T., Martin, N. F., Rix, H.-W., Sand, D. J., Bell, E. F., et al. 2007, The elongated structure of the Hercules Dwarf spheroidal galaxy from deep large binocular telescope imaging. *Astrophys J Lett*, 668, L43–L46
- Davison, W. B. 1990, Design strategies for very large telescopes. *Proc SPIE*, 1236, 878–883
- Davison, W. B., Williams, J. T., & Hill, J. M. 1998, Handling 20 tons of honeycomb mirror with a very gentle touch. *Proc SPIE*, 3352, 216–225
- Davison, W. B., Rascon, M. H., Cuerden, B., Sebag, J., Claver, C., Muller, G., et al. 2004a, LSST structural design. *Proc SPIE*, 5495, 180–188
- Davison, W. B., Warner, S. H., Williams, J. T., Lutz, R. D., Hill, J. M., & Slagle, J. H. 2004b, Handling and transporting the 8.4m mirrors for the large binocular telescope. *Proc SPIE*, 5495, 453–462
- Denman, C. A., Drummond, J. D., Eickhoff, M. L., Fugate, R. Q., Hillman, P. D., Novotny, S. J., et al. 2006, Characteristics of sodium guidestars created by the 50-watt FASOR and first closed-loop AO results at the starfire optical range. *Proc SPIE*, 6272, 67721L
- Dierickx, P., Enard, D., Merkle, F., Noethe, L., & Wilson, R. N. 1990, ESO VLT II: optical specification and performance of large optics. *Proc SPIE*, 1236, 138–151
- Dressler, A., Hare, T., Bigelow, B. C., & Osip, D. J. 2006, IMACS: the wide-field imaging spectrograph on Magellan-Baade. *Proc SPIE*, 6269, 62690F–13
- Esposito, S., Riccardi, A., Quiros-Pacheco, F., Pinna, E., Puglisi, A., Xompero, M., et al. 2010, Laboratory characterization and performance of the high-order adaptive optics system for the large binocular telescope. *Appl Opt*, 49, G174–G189
- Fan, X., Hennawi, J. F., Richards, G. T., Strauss, M. A., Schneider, D. P., Donley, J. L., et al. 2004, A survey of $Z > 5.7$ quasars in the sloan digital sky survey. III. Discovery of five additional quasars. *Astron J*, 128, 515–522
- Fugate, R. Q. 2003, The starfire optical range 3.5-m adaptive optical telescope. *Proc SPIE*, 4837, 934–943
- Goble, L., Angel, J., & Hill, J. M. 1985, Steps toward 8m honeycomb mirrors: VII. Spin casting of an experimental F/1 1.8m honeycomb blank of borosilicate glass. *Proc SPIE*, 571, 92–100
- Goble, L. W., Angel, J., Hill, J. M., & Mannery, E. J. 1989, Spincasting of a 3.5-m diameter f/1.75 mirror blank in borosilicate glass. *Proc SPIE*, 966, 300–308
- Gray, P. M., Hill, J. M., Davison, W. B., Callahan, S. P., & Williams, J. T. 1994, Support of large borosilicate honeycomb mirrors. *Proc SPIE*, 2199, 691–702
- Grazian, A., Menci, N., Giallongi, E., Gallozzi, S., Fontanot, F., Fontana, A., et al. 2009, Wide and deep near-UV (360 nm) galaxy counts and the extragalactic background light with the large binocular camera. *Astron Astrophys*, 505, 1041–1048
- Hastie, M., & McLeod, B. 2008, Comprehensive review of the converted MMT’s instrument suite. *Proc SPIE*, 7014, 70140B–14
- Hénault, F. 2010, PSF and field of view characteristics of imaging and nulling interferometers. *Proc SPIE*, 7734, 773419–14

- Herbert-Fort, S., Zaritsky, D., Christlein, D., & Kannappan, S. J. 2010, The surface mass density and structure of the outer disk of NGC 628. *Astrophys J*, 715, 902–907
- Herbst, T. M., Ragazzoni, R., Eckart, A., & Weigelt, G. 2010, Imaging beyond the fringe: an update on the LINC-NIRVANA Fizeau interferometer for the LBT. *Proc SPIE*, 7734, 773407–7
- Hill, J. M. 1990, Optical design, error budget and specifications for the columbus project telescope. *Proc SPIE*, 1236, 86–107
- Hill, J. M. 1994, Strategy for interferometry with the large binocular telescope. *Proc SPIE*, 2200, 248–259
- Hill, J. M. 2010, The large binocular telescope. *Appl Opt*, 49, D115–D122
- Hill, J. M., & Angel, J. 1983, Steps toward 8m honeycomb mirror blanks: II. Experiments with waffleplates and honeycomb casting. *Proc SPIE*, 380, 100–110
- Hill, J. M., & Angel, J. 1992, The casting of the 6.5m borosilicate mirror for the MMT conversion. *Progress in Telescope and Instrumentation Technologies* (Garching: ESO), 57–66
- Hill, J. M., Hunten, M. R., Johnson, K. J., Mitchell, D., Schaller, S., & Esterline, R. S. 1990, A control system for spincasting 8-meter borosilicate honeycomb mirrors. *Proc SPIE*, 1235, 486–502
- Hill, J. M., Angel, J., Lutz, R. D., Olbert, B. H., & Strittmatter, P. A. 1998, Casting the first 8.4 meter borosilicate honeycomb mirror for the large binocular telescope. *Proc SPIE*, 3352, 172–181
- Hill, J. M., Green, R. F., Slagle, J. H., Ashby, D. S., Brusa-Zappellini, G., Brynnel, J. G., et al. 2008a, The large binocular telescope. *Proc SPIE*, 7012, 701203–15
- Hill, J. M., Ragazzoni, R., Bartuffolo, A., Biddick, C. J., Kuhn, O. P., Diolaiti, E., et al. 2008b, Prime focus active optics with the large binocular telescope. *Proc SPIE*, 7012, 7012M–10
- Hill, J. M., Green, R. F., Ashby, D. S., Brynnel, J. G., Cushing, N. J., Little, J., et al. 2010, The large binocular telescope. *Proc SPIE*, 7733, 77330C–11
- Hinz, P. M., Bippert-Plymate, T., Breuninger, A., Connors, T., Duffy, B., Esposito, S., et al. 2008, Status of the LBT interferometer. *Proc SPIE*, 7013, 701328–9
- Jacobs, S. D., Golini, D., Hsu, Y., Puchebner, B. E., Strafford, D., Prokhorov, I. V., et al. 1995, Magnetorheological finishing: a deterministic process for optics manufacturing. *Proc SPIE*, 2576, 372–382
- Kaiser, N., Aussel, H., Burke, B., Boesgaard, H., Chambers, K., Chun, M. R., et al. 2002, Pan-STARRS: a large synoptic survey telescope array. *Proc SPIE*, 4836, 154–164
- Kantor, J., & Axelrod, T. 2010, The large synoptic survey telescope data management overview. *Proc SPIE*, 7740, 77401N–8
- Keck observatory project office. 1985, Telescope performance specifications. in *The Design of the Keck Observatory and Telescope* (Ten Meter Telescope), ed. J. E. Nelson, T. S. Mast, & S. M. Faber (Berkeley: LBL), 3–1 to 3–13
- Kim, D. W., & Burge, J. H. 2010, Rigid conformal polishing tool using non-linear visco-elastic effect. *Opt Express*, 18, 2242–2257
- Krabendam, V. L., & Sweeney, D. 2010, The large synoptic survey telescope preliminary design overview. *Proc SPIE*, 7733, 77330D–11
- Liu, W. M., Hinz, P. M., Meyer, M. R., Mamajek, E. E., Hoffmann, W. F., Brusa, G., et al. 2007, Observations of Herbig Ae disks with nulling interferometry. *Astrophys J*, 658, 1164–1172
- Lloyd-Hart, M. 2000, Thermal performance enhancement of adaptive optics by use of a deformable secondary mirror. *Pub ASP*, 112, 264–272
- Lubliner, J., & Nelson, J. E. 1980, Stressed mirror polishing: a technique for producing non-axisymmetric mirrors. *App Opt*, 19, 2332–2340
- Mannery, E. J., Siegmund, W. A., & Hull, C. L. 1989, The performance of the apache point observatory 3.5m telescope. *Astrophys Space Sci*, 160, 269–274
- Martin, H. M., Angel, J. R. P., & Cheng, A. Y. 1988, Use of an actively stressed lap to polish a 1.8-m F/1 paraboloid. in *Very Large Telescopes and Their Instrumentation*, ed. M.-H. Ulrich (Garching: ESO), 353–361
- Martin, H. M., Anderson, D. S., Angel, J. R. P., Burge, J. H., Davison, W. B., DeRigne, S. T., et al. 1992, Stressed-Lap polishing of 1.8-m f/1 and 3.5-m f/1.5 primary mirrors. in *Progress in Telescope and Instrumentation Technologies*, ed. M.-H. Ulrich (Garching: ESO), 169–172
- Martin, H. M., Davison, W. B., DeRigne, S. T., Hill, J. M., Hille, B. B., & Trebisky, T. T. 1994, Active supports and force optimization for a 3.5m honeycomb sandwich mirror. *Proc SPIE*, 2199, 251–262
- Martin, H. M., Allen, R. G., Angel, J. R. P., Burge, J. H., Davison, W. B., DeRigne, S. T., et al. 1998a, Fabrication and measured quality of the MMT primary mirror. *Proc SPIE*, 3352, 194–204
- Martin, H. M., Callahan, S. P., Cuerden, B., Davison, W. B., DeRigne, S. T., Dettmann, L. R., et al. 1998b, Active supports and force optimization for the MMT primary mirror. *Proc SPIE*, 3352, 412–423

- Martin, H. M., Allen, R. G., Cuerden, B., DeRigne, S. T., Dettmann, L. R., Ketelsen, D. A., et al. 2000, Primary mirror system for the first magellan telescope. *Proc SPIE*, 4003, 2–13
- Martin, H. M., Cuerden, B., Dettmann, L. D., & Hill, J. M. 2004, Active optics and force optimization for the first 8.4m LBT mirror. *Proc SPIE*, 5489, 826–837
- Martin, H. M., Allen, R. G., Cuerden, B., Hill, J. M., Ketelsen, D. A., Miller, S. M., et al. 2006, Manufacture of the second 8.4m primary mirror for the large binocular telescope. *Proc SPIE*, 6273, 62730C–10
- Martin, H. M., Burge, J. H., Cuerden, B., Davison, W. B., Kingsley, J. S., Lutz, R. D., et al. 2008, Manufacture of a combined primary and tertiary mirror for the large synoptic survey telescope. *Proc SPIE*, 7018, 70180G–12
- Martin, H. M., Allen, R. G., Burge, J. H., Kim, D. W., Kingsley, J. S., Tuell, M. T., et al. 2010, Fabrication and testing of the first 8.4-m off-axis segment for the giant magellan telescope. *Proc SPIE*, 7739, 77390A–13
- McCaughey, G. V. 1934, Making the glass disk for a 200-inch reflecting telescope. *Sci Mon*, 39, 79–86
- Meinel, A. B. 1979, Cost scaling laws applicable to very large optical telescopes. *Proc SPIE*, 172, 2–7
- Miller, N. A., Oegerle, W. R., & Hill, J. M. 2004, The dynamics of abell 2125. *Astrophys J*, 613, 841–850
- Milton, N. M., Lloyd-Hart, M., Baranec, C., Stalcup, T., Powell, K., McCarthy, D., et al. 2008, Commissioning the MMT ground-layer and laser tomography adaptive optics systems. *Proc SPIE*, 7015, 701522–11
- Miroshnikov, M. M., Ljubarsky, S. V., & Khimich, Y. P. 1992, Mirrors for optical telescopes. *Opt Eng*, 31, 701–710
- Morey, G. W. 1938, *The Properties of Glass* (New York: Reinhold)
- Noethe, L. 2002, Active optics in modern, large optical telescopes. *Progress in Optics* (Amsterdam: North-Holland), 3–69
- Olbert, B. H., & Schenck, S. R. 1997, Fracture of an aluminosilicate fiberboard. *J Am Ceram Soc*, 80, 2789–2797
- Olbert, B., Angel, J., Hill, J. M., & Hinman, S. F. 1994, Casting 6.5 meter mirrors for the MMT conversion and magellan. *Proc SPIE*, 2199, 144–155
- Osip, D. J., Phillips, M. M., Palunas, P., Perez, F., & Leroy, M. 2008, Magellan telescopes operations 2008. *Proc SPIE*, 7016, 701609–10
- Parks, R. E., Wortley, R. W., & Cannon, J. E. 1990, Engineering with light weight mirrors. *Proc SPIE*, 1236, 735–743
- Parodi, G., Hill, J. M., & Salinari, P. 1992, Supporting the 8.4-m honeycomb mirrors of columbus. *Progress in Telescope and Instrumentation Technologies* (Garching: ESO), 301–305
- Parodi, G., Cerra, G. C., Hill, J. M., Davison, W. B., & Salinari, P. 1997, LBT primary mirrors: the final design of the supporting system. *Proc SPIE*, 2871, 352–359
- Pease, F. G. 1926, On the design of very large telescopes. *Pub Astron Soc Pac*, 38, 195–207
- Preston, F. W. 1927, The theory and design of plate glass polishing machine. *J Soc Glass Technol*, 11, 214–256
- Prokhorov, I. V., Kordonsky, W. I., Gleb, L. K., Gorodkin, G. R., & Levin, M. L. 1992, New high-precision magnetorheological instrument-based method of polishing optics. *OSA OF&T Workshop Digest*, 134–136
- Rabien, S., Ageorges, N., Barl, L., Beckmann, U., Blumchen, T., Bonaglia, M., et al. 2010, ARGOS: the laser guide star system for the LBT. *Proc SPIE*, 7736, 77360E–12
- Ragazzoni, R., Giallongo, E., Pasian, F., Baruffolo, A., Bertram, R., Diolaiti, E., et al. 2006, The wide-field eyes of the large binocular telescope. *Proc SPIE Int Soc Opt Eng*, 6267, 626710–8
- Ritchey, G. W. 1928, The modern photographic telescope and the new astronomical photography part I. – The fixed universal telescope. *JRASC*, 22, 159–177
- Roddier, N. A., Blanco, D. R., Goble, L. W., & Roddier, C. A. 1995, WIYN telescope active optics system. *Proc SPIE*, 2479, 364–376
- Sabatke, E. E., Burge, J. H., & Hinz, P. 2006, Optical design of interferometric telescopes with wide fields of view. *Appl Opt*, 45, 8026–8035
- Salinari, P., & Hill, J. M. 1994, The enclosure of the large binocular telescope. *Proc SPIE*, 2199, 442–451
- Schechter, P. L., Burley, G. S., Hull, C. L., Johns, M., Martin, H. M., Schaller, S., et al. 2003, Active optics on the baade 6.5-m (Magellan I) telescope. *Proc SPIE*, 4837, 619–627
- Shectman, S. A., & Johns, M. 2003, The Magellan telescopes. *Proc SPIE*, 4837, 910–918
- Shectman, S. A., & Johns, M. 2010, GMT overview. *Proc SPIE*, 7733, 77331Y–11
- Siegmund, W. A., Stepp, L., & Lauroesch, J. 1990, Temperature control of large honeycomb mirrors. *Proc SPIE*, 1236, 834–843
- Storm, J., Seifert, W., Bauer, S. M., Diones, F., Fechner, T., Kraemer, F., et al. 2004, The acquisition, guiding and wavefront error sensing units for the Large Binocular Telescope. *Proc SPIE*, 5489, 374–381

- Su, P., Burge, J. H., Cuerden, B., Sasian, J., & Martin, H. M. 2008, Scanning pentaprism measurements of off-axis aspherics. *Proc SPIE*, 7018, 70183T–10
- Su, P., Burge, J. H., Cuerden, B., Allen, R., & Martin, H. M. 2009, Scanning pentaprism measurements of off-axis aspherics II. *Proc SPIE*, 7426, 74260Y–9
- Teran U., J., Slagle, J. H., Hill, J. M., & Neff, D. 2002, Completion of the large binocular telescope enclosure. *Proc SPIE*, 4837, 217–224
- Tyson, J. A., Wittman, D. M., & Angel, J. R. P. 2001, The dark matter telescope. in *Gravitational Lensing: Recent Progress and Future Goals*, Vol. 237, ed. T. S. Brainerd, & C. S. Kochanek (San Francisco: Astronomical Society of the Pacific), 417–420
- Uglesich, R. R., Crotts, A. P., Baltz, E. A., de Jong, J., Boyle, R. P., & Corbally, C. J. 2004, Evidence of halo microlensing in M31. *Astrophys J*, 612, 877–893
- West, S. C., Callahan, S., Chaffee, F. H., Davison, W., DeRigne, S., Fabricant, D., et al. 1997, Toward first light for the 6.5-m MMT telescope. *Proc SPIE*, 2871, 38–48
- West, S. C., Burge, J. H., Cuerden, B., Davison, W., Hagen, J., Martin, H. M., et al. 2010, Alignment and use of the optical test for the 8.4-m off-axis primary mirrors of the Giant Magellan telescope. *Proc SPIE*, 77390N–15
- Wilson, S. R., & McNeil, J. R. 1987, Neutral ion beam figuring of large optical surfaces. *Proc SPIE*, 818, 320–324
- Woolf, N. J. 1979, Dome seeing. *PASP*, 91, 523–529
- Woolf, N. J. 1982a, High resolution imaging from the ground. *ARAA*, 20, 367–398
- Woolf, N. J. 1982b, Seeing and the design and location of a 15-meter telescope. *Proc SPIE*, 332, 193–197
- York, D. G., Adelman, J., Anderson, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., et al. 2000, The sloan digital sky survey: technical summary. *Astron J*, 120, 1579–1587
- Zanotto, E. D. 1998, Do cathedral glasses flow? *Am J Phys*, 66, 392–399
- Zobrist, T. L. 2009, Application of laser tracker technology for measuring optical surfaces. University of Arizona, Ph.D. dissertation, College of Optical Sciences
- Zobrist, T. L., Burge, J. H., & Martin, H. M. 2010, Accuracy of laser tracker measurements of the GMT 8.4 m off-axis mirror segments. *Proc SPIE*, 7739, 77390S–12

5 Active Thin-Mirror Telescopes

Lothar Noethe · Ray Wilson

European Southern Observatory ESO, Garching, Germany

1	<i>Introduction</i>	187
2	<i>Error Sources and Generated Wavefront Aberrations</i>	188
2.1	Error Sources	188
2.2	Definition of the Telescope Image Quality	190
2.3	Generated Wavefront Aberrations	191
2.3.1	Description of Aberrations	191
2.3.2	Aberrations of a Perfect Telescope	192
2.3.3	Optical Manufacturing	193
2.3.4	Tracking and Misalignments	194
2.3.5	Mirror Deformations	194
2.3.6	Local Air and Free Atmosphere	197
2.3.7	Comparison of Aberrations	198
3	<i>Passive Stabilizing Features</i>	200
3.1	Definition	200
3.2	Telescope Structure	201
3.3	Mirror Geometries and Substrates	201
3.4	Mirror Supports	202
3.4.1	Continuous Versus Point Supports	202
3.4.2	Position-Based Supports	202
3.4.3	Force-Based Astatic Supports	203
3.4.4	Optimal Support Patterns	204
3.5	Limitations of Passive Telescopes	210
4	<i>Active Optics in Telescopes with Meniscus Mirrors</i>	212
4.1	Advantages of Active Optics	212
4.2	Control Strategy	213
4.3	Optical Wavefront Sensors	215
4.3.1	Choice of the Type of Wavefront Sensor	215
4.3.2	Indistinguishable Error Sources	218
4.3.3	Disturbances by the Atmosphere	220
4.4	Correction of Errors	221
4.4.1	Correction of Misalignments	221

4.4.2	Correction of Mirror Shapes	221
4.4.3	Correction of Deflections of the Structure	224
4.5	Historical Development of Active Optics	224
5	<i>Active Telescopes</i>	225
5.1	Telescopes of the Four-Meter Class	225
5.2	Telescopes of the Eight-Meter Class	228
5.2.1	Design of the VLT Active Optics System	228
5.2.2	Open-Loop Performance of the VLT	231
5.2.3	Closed-Loop Performance of the VLT	232
5.2.4	Other Large Active Telescopes	235
5.3	Future Extremely Large Telescopes	236
	<i>Acknowledgments</i>	238
	<i>References</i>	239

Abstract: Until to the late 1980s, large optical telescopes relied on sufficient stiffness of their structures and their mirrors as well as on mechanical design features enhancing their stability to achieve an image quality that was only limited by the errors introduced by the free atmosphere. The only degrees of freedom that were controlled were the tracking and the focus. With the development of computers and electronics, it became feasible to correct the optics of a telescope during the operation, a technique called active optics. Wavefront sensors equipped with CCD detectors could measure the optical errors generated by the telescope with speed and precision, and these errors could then be corrected during observations by realigning the mirrors with respect to each other and by modifying their shapes. In particular, the large tolerances for the errors after polishing allowed for the manufacture of large, thin mirrors with diameters of up to 8 m. With active optics, telescopes with mirror diameters of the order of 4 and even 8 m could routinely achieve a seeing-limited performance. Active optics is also an integral component of the next generation of extremely large telescopes, which by necessity are systems with complex control of all optical elements.

1 Introduction

The goal of astronomical telescopes is to deliver a large number photons with high resolution to instruments located at the focal plane. The light-collecting power can rather easily be increased by enlarging the primary mirror. According to the principles of diffraction optics, this would also improve the resolution if the wavefront errors generated by the atmosphere and by the telescope itself remained below the wavelength of the observed light. Keeping the telescope errors small requires that the positions and the shapes of the optical surfaces do not deviate too strongly from their prescriptions. However, this is not easy to achieve for large telescopes that point at different locations in the sky and are operated in environments with changing temperatures and wind conditions.

Before the development of so-called adaptive optics devices, which correct the errors introduced by the atmosphere, the resolution of ground-based telescopes was fundamentally limited by the atmosphere. The smallest diameters of point-source images were typically of the order of half an arcsecond. Therefore, the goal for telescopes was less ambitious, namely, to deliver only a so-called “seeing-limited” image quality. For this performance, it was sufficient that the errors generated by the telescope itself were small compared with the substantial errors introduced by the atmosphere.

In passive telescopes, traditionally only the tracking and, occasionally, the focusing were controlled. Even the much less ambitious goal of seeing-limited performance could only be achieved if the diameter of the primary mirror was not larger than 2–3 m. Most large passive telescopes with mirror diameters of the order of 4 m would only occasionally reach a seeing-limited performance. Despite the use of mechanical principles and features to improve the stability, they suffered from a lack of stiffness both of the structure and of the mirrors, which caused misalignments and deformations of the mirrors.

Real improvements within the 4-m class and the opening up of the option of building telescopes with much larger diameters were achieved only with the introduction of computerized control of the alignment and the shapes of the mirrors. This technique, called active optics, measures the wavefront errors with optical wavefront sensors and corrects the errors preferably with those elements that generated the errors. An important advantage of active optics is that some stringent performance requirements, which often demand expensive engineering solutions, can be relaxed because the telescope errors can be corrected during operation.

Active optics can be operated in open or in closed loop. In open-loop operation, the signals are taken from calibration tables, whereas in closed-loop operation, they are collected continuously. One of the advantages of closed-loop active optics, which became feasible with the development of CCD cameras, is that the optical quality of the telescope is automatically monitored during observations.

With active optics control applying approximately one correction per minute, telescopes with diameters of 4–10 m could routinely reach a seeing-limited performance, even at much lower costs than traditional passive telescopes.

Future extremely large telescopes face additional challenges. First, with the development of adaptive optics, either the telescope alone or the full system consisting of the telescope and the instrument will be required to deliver diffraction-limited rather than seeing-limited images. Despite the additional correction capabilities supplied by adaptive mirrors, this also imposes stringent requirements on the optical quality of the telescope itself.

Second, their large structures are naturally more flexible. The stiffness can be enhanced by reducing the length of the telescope. However, a compact tube requires much steeper primary mirrors, which increases the sensitivity to misalignments due to gravity or wind.

Third, large errors can often no longer be sufficiently corrected by the large elements generating the errors, since the frequencies required for the corrections may interfere with the mechanical eigenfrequencies of the system. The solution is to correct the bulk of the large amplitude errors with the large elements with a restricted bandwidth and the residual errors with elements of much smaller masses with much higher bandwidths. For diffraction-limited and often even seeing-limited performance, modern very large telescopes therefore require a multistage control with correcting elements of different sizes.

► [Section 2.1](#) gives a short overview of the potential error sources in optical telescopes, ► [Sect. 2.2](#) describes the various metrics that are used for the telescope quality, and ► [Sect. 2.3](#) outlines which wavefront aberrations are generated by the various error sources and how they are best described mathematically. ► [Section 3](#) introduces the passive features used in telescopes to supply the necessary stability against the influences of various error sources and the limitations faced by telescopes using only these passive features. ► [Section 4](#) shows how these problems can be overcome by the control of the optics. Finally, ► [Sect. 5](#) presents the designs of existing telescopes with mirror diameters of the order of 4 and 8 m and of future telescopes with diameters of the order of 30 m or more.

Other reviews of the use of active optics in telescopes can be found in Ray (1991), Hubin and Noethe (1993), in [Chapter 3](#) of Wilson (1999), in [Chapter 8](#) of Bely (2002), and in Noethe (2002). A book dedicated to astronomical optics and elasticity theory discusses in great detail the control of optics using various forms of thin mirrors (Lemaitre 2009). Historical reviews of the use of thin elastic mirrors and the evolution toward active optics are given in Wilson (1999) and Noethe (2009).

2 Error Sources and Generated Wavefront Aberrations

2.1 Error Sources

The optical configuration of a telescope is defined by the relative positions and the figures of the mirrors. Even a hypothetical perfect telescope is not free of aberrations in the field. Additional aberrations can be generated by misalignments and figure errors of the optical elements as well

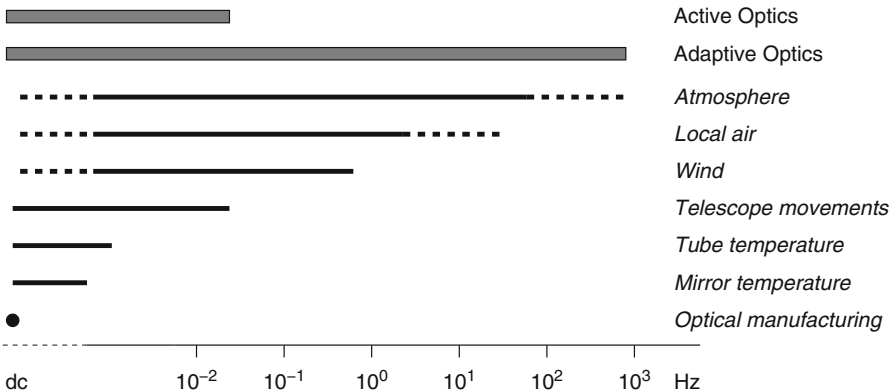


Fig. 5-1

Frequency ranges of sources of wavefront aberrations and of active and adaptive optics corrections in optical telescopes

as by environmental conditions. The frequency ranges of the various error sources discussed below are shown in Fig. 5-1.

Optical Manufacturing. The intrinsic stiffness of large telescope mirrors decreases with their diameter. Their precise shapes are therefore to a large extent defined by their support systems. Under the back surface and along the rim of the mirror, so-called position-based supports define the positions of the support points, whereas so-called force-based supports define the forces at the support points.

On the one hand, the shape of the mirrors is much easier to control with force-based supports, which are therefore the usual choice for the support systems in telescopes. On the other hand, the optical manufacturing process requires comparatively stiff supports if polishing forces are applied on the front surface. However, stiff supports are much harder to adjust than the softer supports used in the telescope during operation, and the difference between the support systems will lead to low spatial frequency errors in the shapes of the mirrors.

Independently of the type of the support system, the polishing process may also generate errors with mid and high spatial frequencies.

Another problem is that in most telescopes, only the combination of all powered mirrors will deliver the required high-quality image. The desired shapes of the individual mirrors usually deviate strongly from spherical shapes, and thus cannot be tested at their centers of curvature with a simple source like an illuminated pinhole generating a perfect spherical wavefront. Instead, the polishing process of single mirrors usually requires special optics, so-called null or compensation systems (Wilson 1999), to convert the spherical wavefront generated by a point-like source into a wavefront with the aspheric shape to be polished into the mirror. Incorrect manufacturing of null systems, which has occurred quite frequently, can generate low spatial frequency wavefront errors of several micrometers.

Telescope Movements. Telescope movements have two effects on the image quality. First, if the telescope does not precisely follow the movement of the stars, the image of a point source will be smeared out because its position moves on the detector. Second, changes of the inclination of the telescope may lead to deformations of the tube and to incorrect mirror support forces, which will generate misalignments and shape errors of the optical elements, respectively.

In addition, the sag of the mirror between discrete support points is usually polished out with the mirror pointing upward. However, some fraction of it will reappear when the axis of the mirror is no longer vertical or, as is the case for secondary mirrors, the mirror is pointing downward.

The frequency range of the generated errors depends strongly on the efficiency of mechanical features that attempt to reduce the misalignments and decouple the deformations of the mirrors from the deformations of their cells. Only small telescopes with mirror diameters up to 2 m are often sufficiently rigid to be unaffected by telescope movements.

Temperature. Temperature variations affect both the telescope structure and the mirrors. The relatively large thermal inertia of the mirrors filters the variations of the ambient temperature, and the generated errors therefore evolve slowly with time.

Wind. In very large telescopes, the wind possibly becomes the most critical source for telescope errors. It applies torques on the main structure, which cause errors in the tracking of the object. In addition, deformations of the structure will lead to misalignments, and pressures on thin mirrors will cause deformations of the mirrors. The errors can be quite large with appreciable contributions even at frequencies of the order of 1 Hz.

Local air. Because of the thermal inertia of the structure and the mirrors, their temperatures cannot follow the fast temperature variations of the ambient air. The temperature differences between the structure and the mirrors on the one hand and the ambient air on the other hand can generate turbulent convection patterns above the surfaces. In the case of the mirror, the impact on the image quality, the so-called the mirror seeing, is approximately proportional to the square of the temperature difference (Guisard et al. 2000; Racine et al. 1991). Therefore, a common requirement is to keep the temperature differences within a range of $\pm 1^\circ\text{C}$.

Free atmosphere. A major error source for ground-based observations are effects caused by atmospheric turbulence that mixes air volumes of different temperatures. The strongest effects are due to the turbulence near the ground and at heights of approximately 10 km. The correction of wavefront errors caused by the free atmosphere is done by adaptive optics systems operating at frequencies up to several hundred Hz.

2.2 Definition of the Telescope Image Quality

Metric for diffraction-limited telescopes. Large telescopes can approach their diffraction limit if they either operate in space or correct the wavefront errors caused by the atmosphere using adaptive optics devices. The common criterion for an image quality close to the diffraction limit is the Strehl ratio s (Born and Wolf 1997), which is the ratio of the peak intensity in the image of a point source taken with the real telescope to the one taken with a corresponding perfect telescope, with both telescopes operating in vacuum. For wavefront errors that are small compared with the wavelength λ of the light, the Strehl ratio depends quadratically on the rms σ of the wavefront error, that is, $s = 1 - (2\pi\sigma/\lambda)^2$ (Born and Wolf 1997). Telescopes are often defined as diffraction-limited if the Strehl ratio is larger than 0.82, which is equivalent to an rms of the wavefront error smaller than a 15th of the wavelength of the observed light.

Metric for the effects of the atmosphere. Without adaptive optics corrections, the image quality even of perfect ground-based telescopes is limited by the wavefront errors generated by the free atmosphere. The metric commonly used for the effects of the atmosphere is called the seeing. It is the full width at half maximum (FWHM) of the image, blurred by the atmosphere, of a perfect large telescope with a diameter of the primary mirror much larger than the atmospheric coherence length or Fried parameter r_0 (Fried 1965). For a seeing of 1 arcsec, r_0 is of the order

of 100 mm. The very best seeing values recorded at the current sites of large telescopes are of the order of 0.3 arcsec FWHM.

Metrics for seeing-limited telescopes. The performance of a telescope is called seeing-limited if the errors introduced by the telescope alone are significantly smaller than the errors introduced by the atmosphere. The wavefront errors due to the atmosphere are so large that also the effects of possibly much smaller telescope errors on the image quality are best described in terms of geometrical optics, for example, by the FWHM of the geometric point spread function generated by the telescope errors.

Without the disturbances of the atmosphere, a large telescope with a seeing-limited performance should generate point spread functions of less than 0.1 arcsec FWHM. Under the assumption that the telescope errors are statistically independent of the errors introduced by the atmosphere, such a figure is quadratically insignificant compared with best seeing values of 0.3 arcsec FWHM.

Design metrics for seeing-limited telescopes. Two other metrics have been introduced for ground-based telescopes. They are based on comparisons of certain characteristics of images taken by the real and a corresponding perfect telescope. Both of these images have to be affected by atmospheric disturbances that generate the same seeing conditions. One of the metrics, the central intensity ratio CIR (Dierickx 1992), is similar to the Strehl ratio. It is the same ratio of the maxima of the images taken by the real and the perfect telescope, this time, however, with both images affected by the atmosphere. For small slope errors the CIR depends quadratically on the rms of the slope errors.

The other metric is the normalized point source sensitivity (PSSN) (Seo et al. 2009), which is defined as the integral of the square of the intensity in the image of the real telescope, normalized by the same integral for the image of the perfect telescope. Again, both images must be identically affected by the atmosphere. The PSSN is directly related to the photometric errors in background limited observations and therefore represents the loss in efficiency in telescope observing time.

For sufficiently small wavefront errors, the last two metrics offer the advantage that the values for the complete system can approximately be computed by multiplying the values corresponding to the individual contributing error sources. The problem with these two metrics is that they depend on the current seeing conditions, which can vary strongly even over short periods of time. Therefore, any of them could only be measured at the telescope if the seeing was well known. However, to define seeing values for the telescope environment measured by independent devices that are close to the telescope has proven to be difficult (Sarazin et al. 2008). These metrics are therefore mainly of theoretical nature, but can conveniently be used for specifications during the design phase.

2.3 Generated Wavefront Aberrations

This section describes which wavefront aberrations are generated by the error sources discussed above, how they scale with the telescope size, and how they are best described mathematically.

2.3.1 Description of Aberrations

Wavefront aberrations $w(\rho, \varphi)$ over a circular pupil, with ρ being the normalized radial and φ the azimuthal cylindrical pupil coordinates, are distributed over a large range of spatial

frequencies. In general, the amplitudes of the aberrations decline rapidly with increasing spatial frequencies. The wavefront error $w(\rho, \varphi)$ is therefore suitable for a modal description in terms of the coefficients of the first terms of its expansion across the pupil in a complete set of preferably orthonormal functions. However, special polishing processes or the geometry of the support system may, in addition, generate characteristic spatial frequencies with possibly large amplitudes.

If the telescope optics is rotationally symmetric, these sets can be expressed as Fourier series. A function $w(\rho, \varphi)$ will then be written as a sum of products of two contributions that depend either on ρ or on φ , that is $w(\rho, \varphi) = \sum c_{m,i} f_{m,i}(\rho) \cos(m\varphi) + s_{m,i} f_{m,i}(\rho) \sin(m\varphi)$, with $c_{m,i}$ and $s_{m,i}$ being the expansion coefficients. The functions $f_{m,i}(\rho)$ form a complete set within each individual rotational symmetry m . The order i of the mode within the rotational symmetry m is related to the number of nodes of the radial function $f_{m,i}(\rho)$.

In general, there exists an infinite number of such sets. Which ones are most appropriate for the description of the wavefront errors depends on the type of the aberrations that are generated by the various error sources.

2.3.2 Aberrations of a Perfect Telescope

Certain wavefront aberrations are naturally generated even by a perfect telescope. In general, the optical quality of a telescope with rotationally symmetric optics can reach its physical limit, that is, the diffraction limit, only along its optical axis in the center of the field. Away from the center, so-called field aberrations occur, which increase with the field angle.

For rotationally symmetric systems, the set of possible field aberrations, together with their dependence on the pupil and field coordinates, can easily be derived from symmetry considerations (Wilson 2004). In such systems, the field aberrations can depend on the coordinate vector $\vec{\rho}$ in the pupil and the field angle vector $\vec{\sigma}$ only through scalars. The only scalars, which can be formed from $\vec{\rho}$ and $\vec{\sigma}$ are ρ^2 , σ^2 , and $\vec{\sigma} \cdot \vec{\rho} = \sigma\rho \cos \varphi$, where φ is the angle between the vectors $\vec{\sigma}$ and $\vec{\rho}$. Consequently, every aberration can be written as a product of functions depending either on the modulus σ of the field angle, the radial pupil coordinate ρ , or the azimuth pupil coordinate φ . Furthermore, any powers of $\cos^m \varphi$ can be rewritten as Fourier series. Finally, with $x = \rho \cos \varphi$ and $y = \rho \sin \varphi$, the pupil dependencies of all field aberrations can also be expressed as polynomials in Cartesian coordinates x and y . The orthonormalized set of these polynomials are the Zernike polynomials (Born and Wolf 1997), which are the functions most commonly used for the description of optical aberrations (Tango 1977).

Using the notation introduced in [Sect. 2.3.1](#), [Table 5-1](#) shows the most important wavefront aberrations expressed as Zernike polynomials. All functions are normalized such that the variance, that is the rms, across the pupil is equal to one.

In general, the coefficients of the field aberration in optical systems decrease rapidly with the rotational symmetry as well as with the order of the Zernike polynomials within each rotational symmetry. For the small field angles of astronomical telescopes, that is, for $\sigma \ll 1$ rad, the dominant aberrations in the field are the ones with the lowest powers of σ . For each rotational symmetry m , the lowest possible power of σ is always equal to m . The dominant field aberrations are, therefore, spherical aberration, which is constant over the field, third-order coma, which depends linearly on the field angle, and third-order astigmatism with a quadratic dependence on the field angle.

■ Table 5-1

Most important wavefront aberrations expressed as Zernike polynomials

Aberration	Rotational symmetry	Order	Symbol	Zernike polynomials
Piston	0	1	$z_{0,1}$	1
Tilt	1	1	$z_{1,1}$	$2\rho \cos \varphi, 2\rho \sin \varphi$
Defocus	0	2	$z_{0,2}$	$\sqrt{3}(2\rho^2 - 1)$
Third-order spherical aberration	0	3	$z_{0,3}$	$\sqrt{5}(6\rho^4 - 6\rho^2 + 1)$
Third-order coma	1	2	$z_{1,2}$	$\sqrt{8}(3\rho^3 - 2\rho) \cos \varphi,$ $\sqrt{8}(3\rho^3 - 2\rho) \sin \varphi$
Third-order astigmatism	2	1	$z_{2,1}$	$\sqrt{6}\rho^2 \cos 2\varphi,$ $\sqrt{6}\rho^2 \sin 2\varphi$

For an arbitrary first-order definition of an optical system, that is, for arbitrary axial positions and radii of curvature of the optical elements, k field aberrations can be eliminated with k aspheric surfaces (Wilson 2004). The aberrations that are usually eliminated first by the optical design are those with the weakest field dependences. Spherical aberration can be eliminated with one parabolic mirror, in addition, linear third-order with two mirrors in the so-called aplanatic Ritchey-Chrétien design, and, furthermore, also quadratic third-order astigmatism with three mirrors in so-called anastigmatic telescopes.

For telescopes with seeing-limited performance and moderate fields of view of the order of 20 arcmin, the Ritchey-Chrétien design is sufficient. However, for telescopes with diffraction-limited performance, the field aberrations even over small fields of view of the order of a few arcminutes have to be minimized. The same applies to large survey telescopes with mirror diameters of up to 8 m and field diameters of possibly a few degrees, in which the field aberrations are usually minimized by a combination of reflective and refractive optics. Therefore, unless the field aberrations are corrected over smaller fields by additional optics in the instruments, diffraction-limited and large survey telescopes require anastigmatic optics with at least three aspheric mirrors.

2.3.3 Optical Manufacturing

Optical manufacturing errors are spread over a large range of spatial frequencies k . The spectrum of the polishing errors is usually rather flat up to a corner frequency and then drops approximately with k^{-3} for spatial frequencies above 7 cycles/m (Hull et al. 2003).

Similar to field aberrations, low spatial frequency errors due to optical manufacturing are most conveniently described by Zernike polynomials. For example, incorrectly manufactured null systems (Wilson 1999) generate predominantly third-order spherical aberration. The high spatial frequency errors are largely independent of the size of the mirror.

All these errors are field independent if the mirror is located in a pupil. Otherwise, the aberrations in the mirror generate additional field aberrations with field dependencies that are different from the ones in rotationally symmetric systems.

2.3.4 Tracking and Misalignments

Tracking errors, which are errors in the pointing of the telescope to the requested position in the sky, manifest themselves as image motions in the focal plane.

Misalignments are caused by relative rigid-body movements of the optical elements, except by rotations around the symmetry axes in the case of circular elements. The generated aberrations can be best described by Zernike polynomials. For example, movements along the optical axis, which preserve the rotational symmetry, generate field-independent defocus and spherical aberration.

Lateral movements or tilts relative to the optical axis break the usual rotational symmetry of the initially aligned system. With respect to the dependence on the pupil coordinates ρ and φ , the generated aberrations are the same as in rotationally symmetric systems, that is, they can be expressed as Zernike polynomials. However, their field dependences are different (Shack and Thompson 1980). In general, the power of the field dependence, that is, the power of σ , of an aberration in the form of a Zernike polynomial generated by a misalignment is at least one order lower than the power of the field dependence of the same Zernike polynomial in the aligned system. For example, whereas the field dependence of third-order astigmatism in the aligned system is quadratic, it can be constant or linear for the additional third-order astigmatism in the misaligned system.

By far the most important additional aberrations generated by lateral displacements or tilts of the mirrors are field-independent third-order coma and third-order astigmatism with a linear field dependence. Furthermore, a misalignment can also generate a tilt of a focal plane, that is, a defocus that is proportional to the difference between the tilted and the ideal focal plane. All these field aberrations in misaligned optical systems can be calculated very elegantly by the method outlined in Shack and Thompson (1980) and Thompson (2005).

The amplitudes of the structural deformations depend on the stiffness of the structure. For similar designs, the eigenfrequencies are roughly inversely proportional to the linear size of the telescope, whereas the deflections due to wind pressures and gravity forces scale with the first and second power of the linear size, respectively.

Another parameter that influences the effects of misalignments on the wavefront error is the focal ratio of the primary mirror. For identical misalignments, the generated third-order coma is proportional to the inverse of the third power of the focal length or, equivalently, the focal ratio of the primary mirror. Focal ratios have evolved from values of about 3 in classical 4 m telescopes to values around and even below 1 in modern large telescopes.

2.3.5 Mirror Deformations

Derivation of elastic modes. The deformations of telescope mirrors generated by external forces are best expanded in elastic modes of the mirrors. For thin shallow shells and if the mirrors are not constrained at the edges, the elastic modes are very similar to the free vibration modes (Noethe 1991) of the mirrors. Elastic modes are deformations with minimum strain energy for a given root mean square of the deformation. The mathematical construction of the minimum-energy modes requires the solution of an eigenvalue problem, which can be formulated as a variational problem

$$\delta(\mathcal{J} - \xi\mathcal{A}) = 0. \quad (5.1)$$

\mathcal{J} is the total elastic energy, \mathcal{A} the rms of the deformation, and ξ a free parameter, which can be interpreted as the energy per unit of the rms of the deflection. The problem is then to minimize the ratio \mathcal{J}/\mathcal{A} among all possible deflections satisfying appropriate boundary conditions. For a thin shallow circular spherica shell, (5.1) leads to a second-order differential equation for the deflections in each rotational symmetry. If the shape of the mirror is not constrained by the supports defining its rigid-body position in space, the appropriate boundary conditions for the usual support systems described in Sect. 3.4 are the ones for free inner and outer edges.

Within each rotational symmetry m , the solutions of the differential equations form a complete set of orthogonal functions, the elastic modes $e_{m,i}$, where the index i denotes the order, as outlined in Sect. 2.3.1. Since, in the case of thin shallow shells, the inertial effects parallel to the surface are negligible, the modes derived from (5.1) are identical to the vibration modes of the shell with angular eigenfrequencies $\omega_{m,i}$. The corresponding eigenvalues $\xi_{m,i}$, which are proportional to the elastic energies of the modes, can then be expressed as

$$\xi_{m,i} = \frac{1}{2} h \gamma \omega_{m,i}^2, \quad (5.2)$$

where γ is the mass density of the mirror and h its thickness. The shapes of the modes can be described analytically as sums of Bessel functions and rational functions ρ^m and ρ^{-m} . They can also be obtained numerically by singular value decomposition of interaction matrices describing deformations at a large number of points across the mirror generated by the application of discrete forces at several points under the mirror.

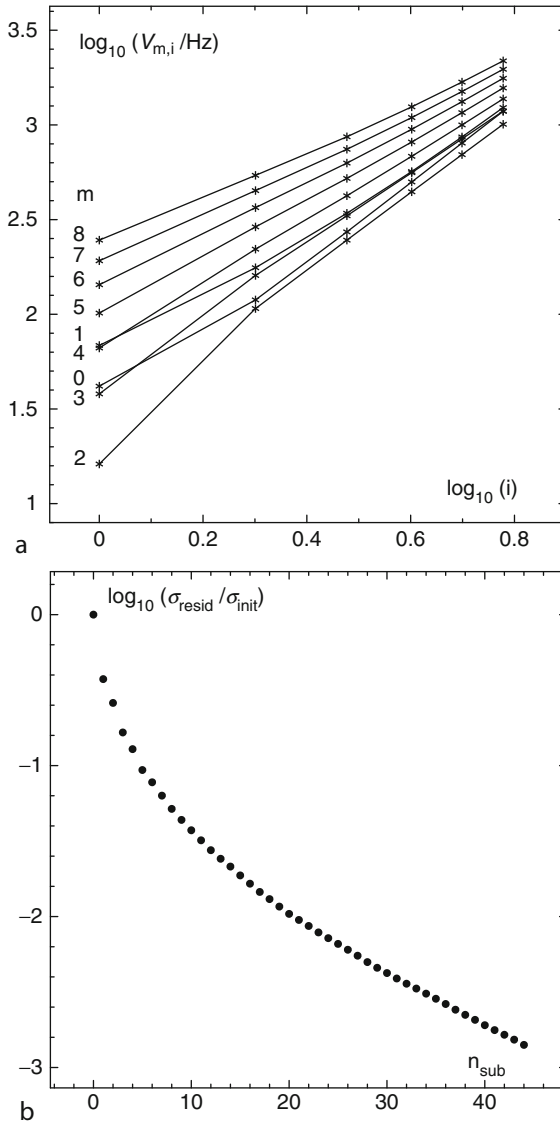
More details about the shapes of the elastic modes are given in Sect. 2.3.7.

Scaling laws. Leaving temperature variations aside, the deformation of a mirror depends essentially on the ratio of external, disturbing forces to the intrinsic stiffness of the mirror. For mirrors with similar geometries and identical substrates, the eigenfrequencies scale with h/D^2 . For a single force acting on a mirror with a diameter D and a thickness h , the deformation is proportional to D^2/h^3 .

Stiffnesses of the modes. In a log-log plot, Fig. 5-2a shows the eigenfrequencies of the lowest elastic modes of a typical thin primary mirror of an 8 m class telescope as functions of their rotational symmetry m and order i within the rotational symmetry. The mirror has a thickness of 175 mm and a radius of curvature of 28.8 m. The most important and useful feature of the elastic modes is that their stiffnesses increase rapidly both with the rotational symmetry m and with the order i . In the log-log plot, the eigenfrequencies increase linearly with slopes of about 2 both with i for constant m and with m for constant i . Therefore, the eigenfrequencies are approximately proportional to $m^2 i^2$.

The strongest deviations from the linear behavior in the log-log plots occur for the lowest modes of the rotational symmetries 0–3. For the rotational symmetries 0 and 1, the relative increase of the stiffnesses is due to strong membrane stresses that arise in thin shells in particular for the lowest modes of these symmetries. The noticeably lower frequencies of the lowest modes of the rotational symmetries 2 and 3 and, to lesser degrees, also 4 are due to their similarity to deformations of the form $\rho^n \cos n\varphi$, which are free of membrane stresses.

The major consequence of the fast increase of the stiffnesses is that higher-order modes containing high spatial frequency deformations are far less likely to be generated than the low-order modes. Therefore, a mirror serves as a filter for the force distribution, reducing very efficiently the effects of forces with high spatial frequency distributions on the deformation of the mirror. Therefore, force errors, which occur naturally in the support of large mirrors, will generate



■ Fig. 5-2

(a) Eigenfrequencies $\nu_{m,i}$ of the elastic modes of the VLT primary mirror for the lowest nine rotational symmetries m and lowest six orders i within each rotational symmetry. (b) Logarithm of the ratio of the rms σ_{resid} of the residual deflections after the subtraction of the n_{sub} lowest elastic modes to the rms σ_{init} of initial deflections generated by random pressure fields

significant amounts of aberrations only in the lowest elastic modes. By far the softest one is the lowest mode $e_{2,1}$ of rotational symmetry 2, which is similar to third-order astigmatism.

► Figure 5-2b shows the logarithm of the statistical average of the ratio $\sigma_{\text{resid}}/\sigma_{\text{init}}$, where σ_{init} is the rms of the initial deflections, which are generated by random white noise pressure

fields, and σ_{resid} is the rms of the residual deflections after the subtraction of the lowest n_{sub} elastic modes. A subtraction of only the softest and therefore dominant mode $e_{2,1}$ reduces the rms of the residual deflection already to 40% and a subtraction of the softest five modes even to 10%.

Calculation of mirror deformations due to applied pressure fields and forces. A pressure field that is proportional to an elastic mode generates deflections exactly in the form of this mode. For all pressure fields in the form of elastic modes, the deflection of the mirror is then proportional to the rms of the pressure field and inversely proportional to the stiffness of this mode.

Arbitrary pressure fields as well as discrete forces, which can be described as delta functions, can be expanded in elastic modes. The amplitudes of the deflections in given modes decrease rapidly with the order of the modes. Except for local effects, the contributions from the higher-order variations in the pressure fields or the force distributions to the deformations are therefore negligible. The low-order deformation of a mirror can then, with high accuracy, be calculated as the linear sum of the low-order contributions from all supports.

Density of support points. The sag of the mirror between neighboring support points creates high spatial frequency aberrations. Together with the thickness of the mirror, the specification for this sag defines the density of the support points.

Mass reduction. One way to reduce the mass or increase the stiffness of monolithic mirrors is to use structured instead of solid mirrors. An even larger mass reduction can be achieved by segmenting the mirror. However, by itself, a segmented mirror has no global stiffness, which must then be supplied by a fast control loop that stabilizes the relative positions of the segments. In this loop, the segment positions are modified by three position actuators under each segment, based on signals from edge sensors, which measure primarily the steps between adjacent segments.

The effects of actuator movements on the signals measured by the edge sensors can be described by an interaction matrix. For a finite number of segments, segment alignment modes are defined as the orthogonal sets of vectors of actuator displacements that generate orthogonal sets of vectors of edge sensor signals. Both sets of vectors can be obtained by a singular value decomposition of the interaction matrix. In the limiting case where the size of the segments goes to zero and at the same time the number of segments to infinity, the segment alignment modes become smooth functions, which are all combinations of Bessel functions (Noethe 2005).

Field aberrations generated by mirror deformations. If a mirror is not in a pupil of the telescope, a deformation of the mirror will generate additional field aberrations. For an aberration described by a Zernike polynomial $z_{m,i}$ with a power $n_\rho = m + 2(i - 1)$ of the radial coordinate ρ , the sums of the powers of ρ and the field coordinate σ of the generated field aberrations are always equal to n_ρ . For example, a deformation of a nonpupil mirror in the form of third-order astigmatism with $n_\rho = 2$ will generate a field aberration that is linear in ρ and σ , which is a field dependent tilt, that is a distortion.

2.3.6 Local Air and Free Atmosphere


The spatial spectrum of wavefront aberrations generated by the free atmosphere is usually described by a von Kármán spectrum, which is constant for small frequencies and declines like a Kolmogoroff spectrum with a power law for frequencies above a certain corner frequency. The wavefront aberrations themselves are best expanded in orthogonal functions that are statistically independent across the pupil, the so-called Karhunen-Loève modes (Roddier 1999).


The description of the aberrations generated by the local air inside the dome is more difficult since it depends on several parameters like the geometries of the dome and the telescope structure, the complex air flow inside the dome, and temperature differences between the elements of the telescope and the air. As a coarse approximation, the spectrum is often described by a von Kármán model with, for example, the corner frequency defined by the size of the opening of the dome and the velocity of the incoming wind.

2.3.7 Comparison of Aberrations

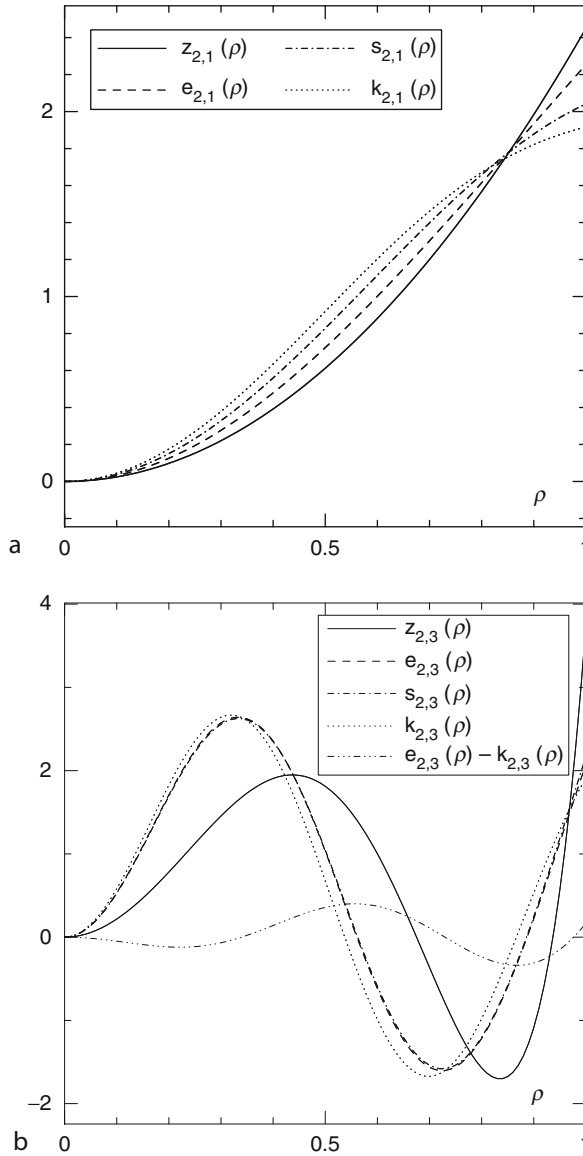
General similarities and differences. All four types of functions or modes introduced in the previous sections to describe the effects of the error sources on the wavefront are qualitatively similar. Being defined on a circular disc, they can be written as products of cosine or sine functions depending only on the polar angle φ on the one hand with functions depending only on the normalized radial coordinate ρ on the other hand. The order i within each rotational symmetry is characterized by the number of nodes along the range of the radial coordinate.

For the rotational symmetries larger than 1, the modes of a given order are qualitatively similar for all four types of modes. However, the first-order modes of the rotational symmetries 0 and 1 are different for the Zernike and Karhunen-Loève modes on the one hand and the elastic and segment alignment modes on the other hand. For the symmetry 0, the lowest Zernike mode is a constant, also called piston, and for the symmetry 1, it is a tilt. Both modes represent rigid-body modes, which do not generate elastic stresses or signals in edge sensors that only measure the steps between adjacent segments. The lowest modes of the elastic and the segment alignment modes of the symmetries 0 and 1 are therefore similar to the second Zernike modes of these symmetries, namely, to defocus and third-order coma, respectively.

Often, edge sensors in segmented mirrors can only measure the steps between two segments, but not the relative tilts around an axis parallel to their adjacent edges. A so-called defocus mode of rotational symmetry 0, which is generated by identical tilts and zero steps between all adjacent segments, is then invisible to the edge sensors. The lowest segment alignment mode of rotational symmetry zero is then not equivalent to Zernike defocus, but is similar to the third Zernike mode of this symmetry, representing third-order spherical aberration. For the rotational symmetry 2,  Figs. 5-3a, b show all four types of modes with the orders 1 and 3, respectively.

Zernike polynomials versus elastic modes. Zernike polynomials are most strongly curved near the outer edge. On the contrary, elastic modes tend to minimize the curvature in the large areas toward the outer edge of the mirror since this minimizes the total strain energy for a given rms of the deformation. Consequently, the elastic modes are effectively linear near the outer edge but more strongly curved near the inner edge than the Zernike polynomials, which can be seen by comparing the solid and the dashed lines in  Fig. 5-3.

The differences increase with the order of the modes. As a consequence, the higher-order elastic modes cannot be well approximated by a small number of annular Zernike polynomials. For example, within the rotational symmetry 2, an elastic mode of the order i can be fitted by the same number i of the lowest modes of the Zernike polynomials. The relative residuals are then only 0.05 for $i = 1$, but 0.35 for $i = 2$ and as large as 0.62 for $i = 3$. Four and six annular Zernike polynomials are required for the fit to push the relative residuals below 0.05 for the orders $i = 2$ and $i = 3$, respectively.



■ Fig. 5-3

Different types of orthonormal modes of rotational symmetry 2. *Solid*: Zernike polynomials $z_{2,i}$, *dashed*: elastic modes $e_{2,i}$, *dashed-dotted*: segment alignment modes $s_{2,i}$, *dotted*: Karhunen-Loève modes $k_{2,i}$. (a) Order $i = 1$, (b) order $i = 3$. Line with three dots in figure b: difference $e_{2,i} - k_{2,i}$ between the elastic and the Karhunen-Loève modes

Nevertheless, viewed as elements of a vector space, at least the lowest elastic modes of the lowest rotational symmetries are nearly parallel to their corresponding Zernike polynomials. Examples of such pairs are Zernike defocus $z_{0,2}$ and the first elastic mode $e_{0,1}$ within the rotational symmetry 0, Zernike third-order coma $z_{1,2}$ and the first elastic mode $e_{1,1}$ within the

rotational symmetry 1, and Zernike third-order astigmatism $z_{2,1}$ and the first elastic mode $e_{2,1}$ within the rotational symmetry 2.

Segmented mirror and Karhunen-Loève modes versus Zernike polynomials. For the higher orders, the segment alignment modes $s_{m,i}$ converge rapidly to the corresponding elastic modes $e_{m,i}$. The reason is that the relative displacements of segments in height measured by two edge sensors per border are equivalent to local torsions in monolithic mirrors. The differential equations, from which the segment alignment and the elastic modes for thin shells are derived, are identical, and the differences between the boundary conditions decrease with the increase of the order of the modes (Noethe 2005).

• *Figure 5-3b* shows that also the higher-order Karhunen-Loève modes $k_{m,i}$ are similar to the corresponding elastic modes $e_{m,i}$. For orders larger than 2, the rms of the difference $k_{m,i}(\rho) - e_{m,i}(\rho)$ between these two types of modes is always of the order of 15%, as shown in • *Fig. 5-3b* by the line with one dash and three dots.

In summary, the higher orders of the elastic, segment alignment and Karhunen-Loève modes are rather similar. They are, however, very different from the corresponding Zernike polynomials. Therefore, except for the lowest-order modes, the Zernike polynomials are not well suited for expansions of the modes of any of the other three sets.

Choice of set of modes for an expansion of the wavefront error. The choice of the set of functions that are used for the expansion of the wavefront error depends on the type of aberrations expected to be generated by the system. The two low-order aberrations Zernike defocus and third-order coma are, in general also with large amplitudes, generated by misalignments and are usually included in the set. The easiest choice is then to use only Zernike polynomials and reconstruct other functions from the coefficients of the Zernike polynomials.

However, the best fit or the smallest residual error will be obtained if the chosen functions best describe the expected aberrations in the wavefront. Since several types of errors sources are usually present in a telescope, the best fit requires the use of different types of functions. In general, these functions will not be orthogonal, and functions from different sets may even be very similar. In such cases, only one function out of a group of similar functions should be included in the set of fitted functions. The same considerations apply if not the functions themselves are used, but instead their derivatives, which could be convenient for those wavefront sensors that measure the slopes of a wavefront.

3 Passive Stabilizing Features

3.1 Definition

Several mechanical features can be used in telescopes to minimize the deleterious effects of some of the error sources on the image quality. Those features that do not require any control are called passive.

Passive telescopes were the standard telescopes until the late 1980s. Apart from occasional refocusing, no corrections of the alignment or the shapes of the mirrors could be done, and, usually, only the tracking of the telescope was controlled. For equatorial mounts, this required only small corrections of the pointing positions generated by a rotation at constant speed around the axis parallel to the axis of the earth.

The largest passive telescopes had diameters of the primary mirror of the order of 5 m. Naturally, they suffered from deformations of the structure and of the mirrors primarily due to

changes of the inclination of the telescope tube. However, passive telescopes profited from constructional mechanical principles for the design of the structure and the mirror supports, which significantly reduced the deformations. In addition, since the correction of atmospheric effects was not possible at that time, they only had to achieve a seeing-limited performance.

3.2 Telescope Structure

One of the most serious sources for wavefront errors is a relative shift or tilt of the secondary with respect to the primary mirror. The structures holding the cells of both mirrors are usually attached via substructures to the centerpiece, the central part of the telescope tube. A particular substructure is the Serrurier truss (Diffrient 1994), which was first used in the 200-in Hale telescope. A proper dimensioning of the struts ensures that the cells of the two mirrors move laterally by the same amount and also without relative tilt. Therefore, the two mirrors remain aligned with respect to each other irrespective of the inclination of the tube.

3.3 Mirror Geometries and Substrates

► *Figure 5-4* shows a typical primary mirror of a passive 4 m class telescope with a thickness of 600 mm at the outer edge and a weight of approximately 12 t. The largest monolithic mirror with a diameter of 6 m, a thickness of 700 mm, and a weight of 45 t is that of the BTA-6 telescope.

One way of increasing the stiffness of a mirror without increasing its mass is lightweighting. The first example of such a structured large mirror was the 200 in mirror of the Hale telescope



■ Fig. 5-4
Primary mirror of the 4.0-m telescope on Cerro Tololo

with approximately 84, mostly triangular, cavities and a lightweighting of approximately 60%. Later on, lightweighting was used for the primary mirror of the Hubble Space Telescope, the 6.5 m mirror of the MMT and the 8.4 m blanks of the Large Binocular Telescope. The latter, with a total thickness at the outer edge of the blank of approximately 900 mm, were cast in a spinning oven. During the casting, the 1,662 honeycombs cavities and the thickness of 11 mm of the ribs were defined by pieces of ceramic. With a lightweighting of 78%, this structured mirror is, in terms of forces required to generate a certain deformation, approximately four times stiffer than the solid monolithic mirrors of the VLT, which have roughly the same diameter and a 50% higher mass.

The shapes of mirrors are also sensitive to temperature variations. The major effect for a powered mirror is a change of the radius of curvature, which generates mainly defocus in the focal plane. Inhomogeneities of the coefficient of thermal expansion within the mirrors may also lead to other deformations in the lowest elastic modes. One solution to prevent these temperature related effects, which has been applied in most modern telescopes, is the use of mirror substrates with negligible coefficients of thermal expansion.

3.4 Mirror Supports

3.4.1 Continuous Versus Point Supports

Lenses can only be supported along their rim, which implies that an essentially one-dimensional support has to maintain the proper three-dimensional shape of a lens under different inclinations. Mirrors, on the other hand, can also be supported at the back surface. They are therefore the only feasible option for large optical elements in large telescopes.

Ideally, axial supports across the back surface and lateral supports along the rim would be continuous. Such supports, which would have to apply pressure fields with possibly special functional dependencies, are difficult to design. Therefore, all large telescope mirrors are supported at discrete points. The additional unwanted local deformations introduced by the point supports die off rather quickly due to the filtering effect of the elastic medium. A proper distribution of the point supports across the back surface and the rim will, to a great extent, also avoid the generation of low spatial frequency errors.

Aberrations with high spatial frequencies generated by the sag between the supports can be limited to tolerable amounts by choosing an adequate density of the support points, once the thickness h and the substrate of the mirror have been defined. If d is the typical distance between neighboring supports, the sag scales with d^4/h^2 .

3.4.2 Position-Based Supports

In the past, mirrors rested on hard support points, which themselves were rigidly attached to the mirror cell. Depending on the errors in the vertical positions of the hard points on the one hand and the intrinsic stiffness of the mirror on the other hand, the shape of the mirror would either follow the support points or the mirror would be supported by only a reduced number of supports. In extreme cases, the mirror could even rest on three points only.

At best, these three points would form a regular triangle and would be located at a distance of approximately 75% of the outer radius from the center. The dominant aberration would then be in the form of the lowest elastic mode $e_{3,1}$ of rotational symmetry 3. For a mirror with a

diameter of 1 m and a thickness of 100 mm, the coefficient of this mode would be of the order 300 nm. Such an error would cause a blur of a spot with a FWHM of the order of 1 arcsec, which would be far above the limit even for a seeing-limited performance.

Since the wavefront errors scale with D^2/h^3 for the effect of a single force and the forces applied by three support points scale with D^2h , the deformation on a three-point support scales with D^4/h^2 . Therefore, larger mirrors would have to be prohibitively thick to reduce the errors to acceptable values.

Even if a mirror was supported by several hard points, the positions of the support points would require accuracies in the axial direction better than 1 μm . However, because of manufacturing errors as well as deformations of the mirror cell, such accuracies are impossible to achieve. Position-based supports are, therefore, ruled out for large monolithic mirrors. A much more practical solution for maintaining the proper figure of the mirror is to use force-based supports, which can, to a great extent, decouple the deformations of the mirror from the deformations of the mirror cell.

3.4.3 Force-Based Astatic Supports

Definition. The shape of a mirror is defined by its shape in a force-free environment and by the forces acting on it: the gravity field, exerting a body force, and the localized support forces. In any of the two extreme positions, that is, when the mirror is either horizontal or vertical, its full weight is supported either by the axial or the lateral support system.

For arbitrary inclination angles ϑ , the gravity forces acting along the axial and lateral directions are proportional to $\cos \vartheta$ and $\sin \vartheta$, respectively. The same dependencies on ϑ must then be valid for the applied axial and lateral support forces as well as for the errors generated by the axial and lateral supports. Therefore, if the axial or lateral forces generate a nearly perfect shape when the mirror is in horizontal or lateral position, they will maintain the proper shape of the mirror for all inclination angles.

Furthermore, the rigid-body position of the mirror with respect to its cell must be defined by so-called fixed points, which constrain the movements in the six potential rigid-body degrees of freedom without affecting the shape of the mirror. A support system with such fixed points, which, in addition, guarantees that the shape of the mirror is decoupled from the shape of the mirror cell, is called astatic.

For maintenance purposes that require the removal of the mirror from its cell, it is an advantage not to have fixed connections between the mirror and the support points. However, such so-called push-only supports cannot apply pull forces on the mirror. While it is possible to achieve a near-perfect mirror shape with push-only axial supports for the mirror in horizontal position, this is not usually the case with push-only lateral supports for a mirror in vertical position.

Lever-type astatic supports. Two frequently used types of astatic supports use mechanical levers. One of them is the whiffle tree support (Bely 2002), which, possibly using more than one stage, distributes a force over several support points at the back surface of the mirror. At each stage, the support force is astatically distributed via a pivoting lever to two or three points in the next level. The fractions of the forces applied by the individual support points therefore depend only on the geometry of the whiffle tree. This type of support is the preferred choice for mirrors with diameters up to approximately 2 m. Larger thin mirrors need a large number of support points, which would inevitably require complicated whiffle trees with several stages.

A better solution is then to use astatic lever supports. Here, the support point at one end of the lever, the rotation axis of the lever, and the center of gravity of the counterweight at the other

end of the lever are located on one line. In such a configuration, the applied individual forces are independent of the inclinations of their levers, which, in turn, depend on the local distances between the mirror and its cell. Consequently, the applied individual forces are decoupled from the deformations of the mirror cell (Lassell 1842) and, therefore, astatic.

In an axial whiffle tree support, three fixed points are required to constrain the three rigid-body degrees of freedom: a translation perpendicular to the surface, called a piston, and rotations around two orthogonal axes parallel to the surface, called tip and tilt. The fixed points are realized by using three whiffle trees, with the bases of the whiffle trees being attached to the cell at angles of 120° .

In astatic lever supports, three points could be added, at which the mirror is coupled to its cell and which should ideally carry no weight. Another solution is to convert three of the astatic supports into fixed points. Since the axial positions of the fixed points do not affect the mirror shape and since all the other individual supports are astatic, also the full support is astatic.

Hydraulic and pneumatic supports. An alternative to lever-type supports, which add an appreciable amount of weight to the mirror cell, are hydraulic or pneumatic supports. Here, a fixed point is defined by a constant volume in a single support or in a group of connected supports. Therefore, the number of groups of connected supports must not be larger than the number of degrees of freedom for the rigid-body movements. As in the lever-type supports, piston, tip, and tilt of a mirror are defined by three fixed points, that is, by three sectors of connected axial supports, usually following a threefold symmetry.

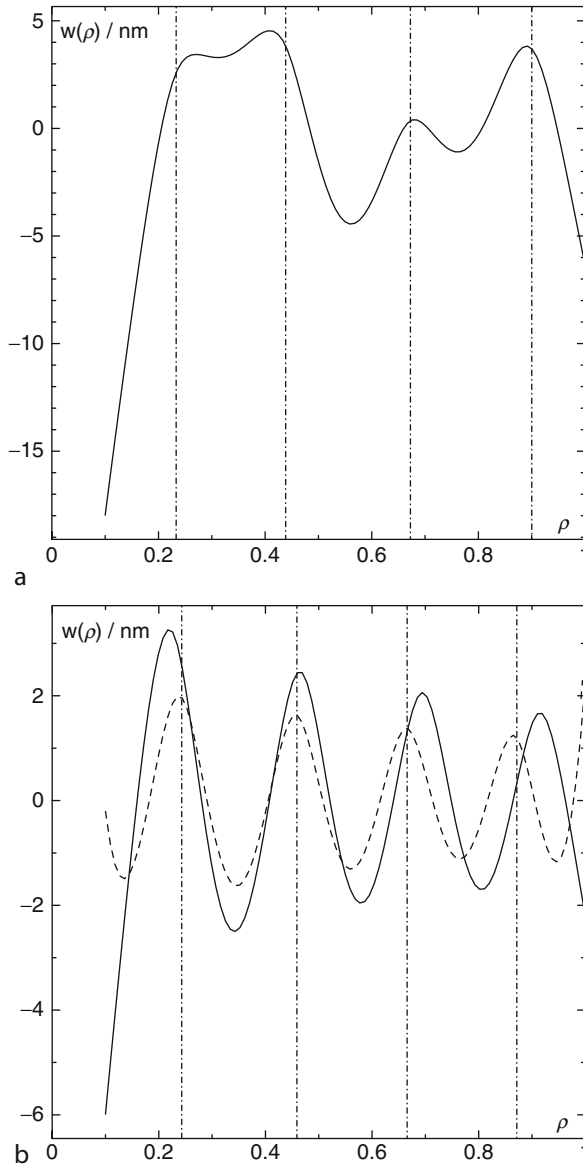
Since the liquid or the air can freely flow within one sector, the forces are not affected by deformations of the cell. Therefore, hydraulic and pneumatic supports designed this way fulfill the two conditions for the astaticity of a support system.

One disturbing feature of these connected supports is that in nonhorizontal positions of the mirror, the hydrostatic pressure produces additional forces, which increase from the highest to the lowest axial support point. In pneumatic supports, such additional pressures are clearly negligible, whereas in hydraulic supports, a two-chamber system must be used to compensate the additional forces (Stanghellini et al. 1997).

3.4.4 Optimal Support Patterns


The patterns and the force distributions of the axial and lateral supports have to be chosen such that the mirror figure is close to the prescribed figure for any zenith angle, in particular when the mirror surface is horizontal or vertical. Since in these limiting cases the full weight is either supported by the axial or the lateral supports, they can be designed separately. An advantage of thin solid meniscus mirrors is that they can be approximated mathematically as two-dimensional entities. Many problems related to the mirror support can then be treated analytically.

Axial supports. The optimum pattern to minimize the rms of the surface error with a given number of supports should be a regular pattern consisting of equilateral triangles. However, for circular mirrors this leads to rather irregular sags at the outer edge. Therefore, the traditional choice is a pattern with supports on concentric rings. Only two parameters are available for the optimization of such supports: the radii of the rings and the fractions of the load supported by each ring. The optimization is a nonlinear process. However, since analytical formulas exist for thin meniscus mirrors (Schwesinger 1988), it can be done rapidly and efficiently by trial and error methods.




■ Fig. 5-5

Minimum wavefront errors $w(\rho)$ generated by a thin plate supported on four continuous rings. (a) Load fractions defined, only free parameters: radii of the rings. (b) *Solid line*: Free parameters: load fractions and radii of the rings, *Dashed line*: Additional free parameter: defocus

In particular for hydraulic or pneumatic supports, it can be an advantage if all supports apply the same force. If, in addition, the number of supports on each ring is defined, the fractions of the total load supported by each ring are fixed. Therefore, the radii of the rings are the only free parameters for minimizing the rms of the deformation. As an example,  Fig. 5-5a shows


the optimized radii and the corresponding radial deformation for a support with four rings. The mirror has an outer diameter of 2 m, a diameter of the inner hole of 200 mm, a radius of curvature of 10 m, and a thickness of 100 mm. It is supported by 6, 12, 18, and 24 forces on four rings. Clearly, such a support cannot generate a deformation with similar sags between the rings. Also, a different distribution of the total number of approximately 60 supports on the rings will not significantly improve the rms of the deformation.

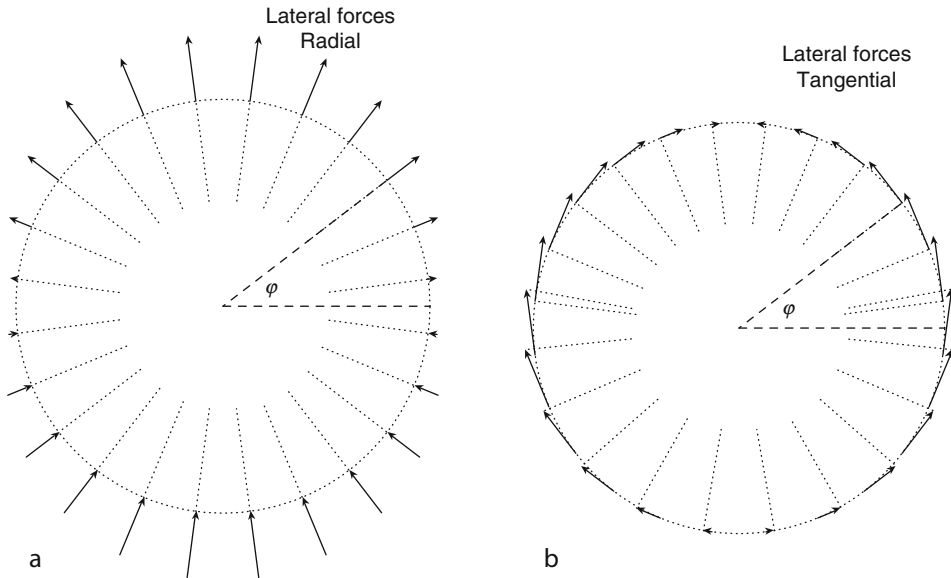
A homogeneous pattern can only be obtained if also the other degrees of freedom, that is, the load fractions, are used.  Figure 5-5b shows that in that case the sags between the rings decrease from the center to the outer edge because the areas of annular sections with the same diameter increase toward the outer edge. Furthermore, a certain amount of defocus can be tolerated since it can be corrected by movements of mirrors along the optical axis. Using also this additional degree of freedom, the rms of the wavefront error of 1.6 nm can be reduced to 1.0 nm with a defocus of $-1.06 \mu\text{m } \rho^2$. Finally, the number of supports on each ring is chosen such that the sags between the neighboring supports on one ring are approximately equal to the sags between neighboring rings.

Scaling laws for axial supports. The sag between adjacent support points scales with d^4/h^2 , where d is the distance between the support points. With D denoting the diameter of the mirror and n_{supp} the number of support points, $d \propto D/\sqrt{n_{\text{supp}}}$. Therefore, the number n_{sag} of supports that are required to limit the sag to a given value scales with D^2/h and the nominal support forces scale with h^2 . With the assumption that the errors in the applied passive forces are proportional to the nominal forces and using that the deflections for the application of a single force scale with D^2/h^3 , the total low spatial frequency wavefront errors w_p generated by n_{sag} random forces under a passive mirror scale with $w_p \propto \sqrt{D^2/h} \cdot h^2 \cdot D^2/h^3 = (D^2/h)^{3/2}$.

Lateral supports. Lateral and axial supports may either be combined or separated units. In combined units, the individual supports apply both axial and lateral forces. If the lateral forces are not applied in the neutral surface of the mirror halfway between the back and front surface, they generate additional unwanted moments. In passive telescopes, the attachment points should therefore be in the neutral surface. This can rather easily be done in structured mirrors with sufficiently thick ribs, but requires additional bores from the back surface up to the neutral surface in solid mirrors.

If the lateral support is independent of the axial one, the support points are usually only located at the outer rim. If the meniscus mirrors are neither too large nor too steep, the plane perpendicular to the axis of the mirror and containing its center of gravity intersects the rim. The mirror can then be supported along the rim under its center of gravity with all force vectors lying in this plane. However, for large, thin, and strongly curved mirrors, the plane through the center of gravity does not intersect the rim. A lateral support system with forces all lying in one plane and acting on the rim would then generate a torque (Schwesinger 1994). To balance this torque, the lateral forces must also contain components along the axis of the mirror.

Analytical expressions can be derived for the deformations of thin meniscus mirrors due to lateral support forces (Schwesinger 1988, 1991). The lateral forces are expressed as Fourier components depending on the azimuth angle φ , which is measured with respect to the horizontal axis ( Fig. 5-6). Only forces proportional to the sine or the cosine of φ can support any weight.



■ Fig. 5-6
Lateral supports with purely (a) radial and (b) tangential forces

Consequently, the lateral support should only apply forces that are proportional to $\sin \varphi$ and $\cos \varphi$. Since such forces only generate deformations that are also proportional to $\sin \varphi$ and $\cos \varphi$, the major task of the design of the lateral support system is to minimize the deformations in the rotational symmetry 1.

The vectors of the forces applied at the rim can be split into three orthogonal components: a radial, a tangential, and an axial one. The radial components will be proportional to $\sin \varphi$ and the tangential ones to $\cos \varphi$, as shown in ► Fig. 5-6. The axial forces, which are required to balance the torque when the mirror is supported at the center of the rim, are proportional to $\sin \varphi$.

With the functional dependence on φ given, the only parameters that need to be defined are the maxima of the three force components. The maximum of the axial forces is determined by the requirement to balance the torque. The sum of the maxima of the radial and tangential forces is defined by the requirement to support the full weight of the mirror. Therefore, the only free parameter for the definition of the lateral support is the ratio of the maxima of the radial to the tangential forces or, in other words, which fraction β of the weight is supported by the tangential forces.

Lateral supports of the NTT and the VLT. An example for a lateral support without axial components is the one of the primary mirror of the ESO New Technology Telescope (NTT). This mirror has a diameter of 3.58 m, a thickness of 241 mm, and a moderate focal ratio of 2.2. It can laterally be supported in the plane through its center of gravity with, therefore, no need for axial force components. The wavefront error does not depend critically on the ratio β of the fraction of the weight supported by tangential forces (► Fig. 5-7a). With the chosen value of $\beta = 0.5$, the

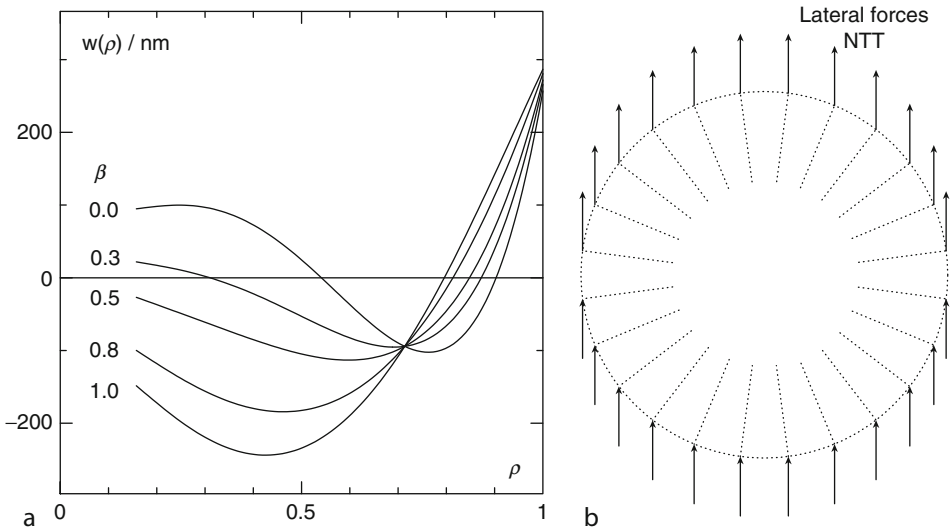


Fig. 5-7

(a) Wavefront error $w(\rho)$ generated by the NTT primary mirror along a normalized radius ρ in vertical direction due to the lateral supports for various fractions β of the weight supported by the tangential forces. (b) Lateral support forces for $\beta = 0.5$

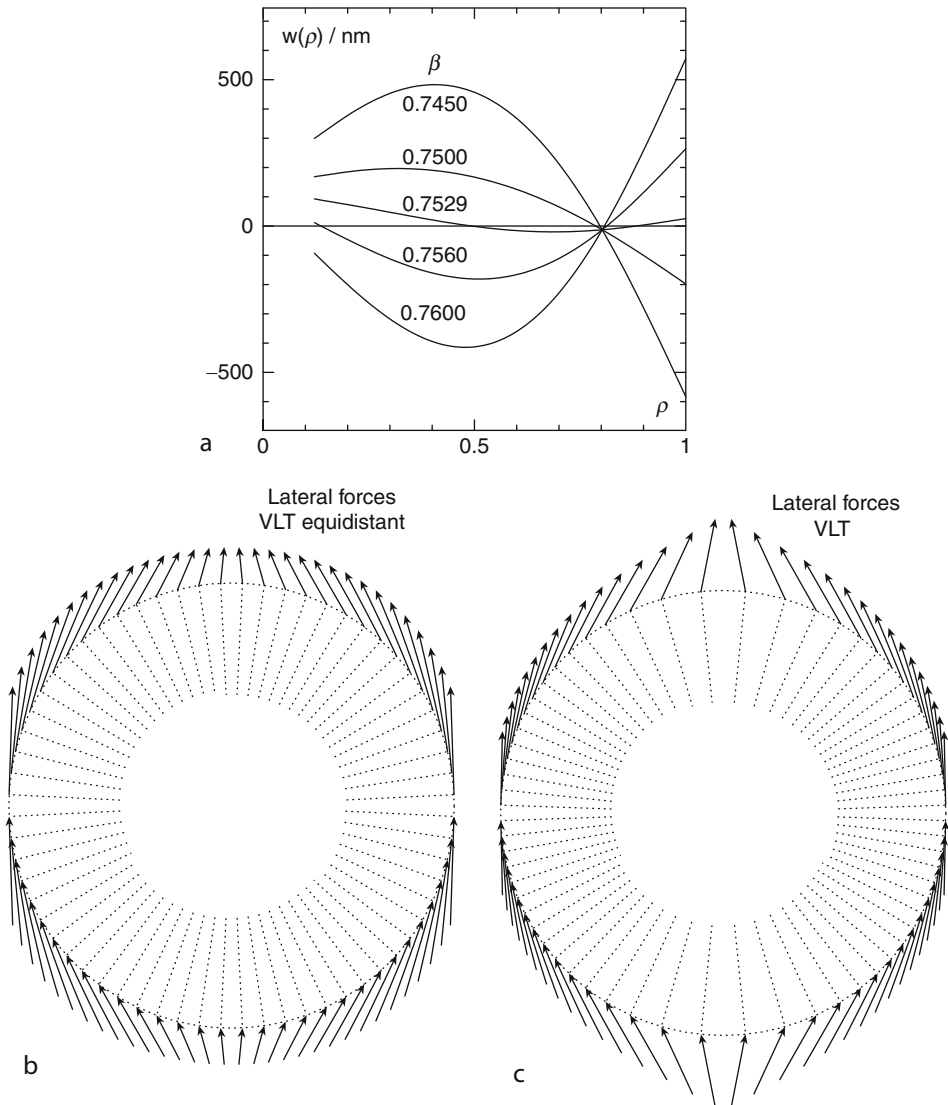
equidistant lateral forces all point upward and are all identical (Fig. 5-7b), which considerably simplified the mechanics.

On the contrary, the large mirrors of the ESO Very Large Telescope (VLT) with diameters of 8 m, thicknesses of 175 mm, and f-ratios of 1.75 cannot be supported in the plane containing their centers of gravity. Since the lateral forces are applied at the center of the rim, they must also have components in the axial direction. The vectorial sums of the axial and the radial components are approximately parallel to the slope of the mirror at the outer edge.

The wavefront error is very sensitive to the fraction β , with the minimum with a rms of 8.7 nm obtained for $\beta = 0.7529$ (Fig. 5-8a). Additional supports at the inner edge or attachments of the supports away from the center of the rim do not yield any improvements (Schwesinger 1991). The better quality achieved for the much more flexible VLT as compared with the one for the NTT is due to the additional axial forces, which were not used for the NTT.

A lateral support system with equidistant support points and 75% of the load carried by tangential forces would lead to forces three times larger at the two sides of the mirror than at the top and the bottom (Fig. 5-8b). However, forces with the same moduli can be applied if the density of the support points increases towards the sides of the mirror (Fig. 5-8c).

The analytical method used for calculating deflections generated by forces with the rotational symmetry 1 can be extended to arbitrary rotational symmetries (Noethe 2001). Densely applied lateral forces that are proportional to $\cos m\varphi$ or $\sin m\varphi$ will only generate deflections with the rotational symmetry m . Similar to the case of axial supports, this feature offers a convenient and fast way of calculating deflections generated by arbitrary forces along the rim in a modal fashion in terms of rotational symmetries.



■ Fig. 5-8

(a) Wavefront error $w(\rho)$ generated by the VLT primary mirror along a normalized radius ρ in vertical direction due to the lateral supports for various ratios β of the weight supported by tangential forces. (b) and (c) Lateral support forces projected onto the plane perpendicular to the axis of the mirror for $\beta = 0.75$ with equidistantly distributed support points (b) and identical moduli of the forces but varying densities of the attachment points along the rim (c)

3.5 Limitations of Passive Telescopes

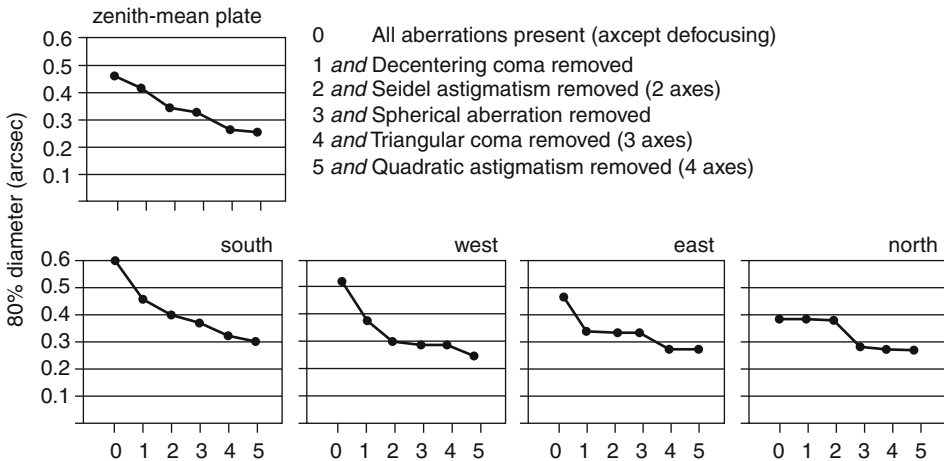
Performance of a passive 4-m telescope. Despite the use of stabilizing constructional mechanical principles, the image quality of large passive telescopes suffered inevitably from deformations of the structure and the mirrors. Wavefront aberrations generated in passive telescopes were investigated systematically at the equatorial ESO 3.6 m telescope on La Silla (Wilson 1982). A Shack-Hartmann sensor measured the coefficients of the lowest-order Zernike polynomials with the telescope pointing at different sky locations.

After the mathematical subtraction of the effects of the low-order aberrations, the image quality was always of the order of 0.25–0.30 arcsec FWHM, independent of the sky location (► Fig. 5-9). These values can be regarded as the intrinsic quality of the telescope, which is defined as a quality obtained without the wavefront aberrations generated by deflections of the structure and errors in the mirror support. Under average seeing conditions of 0.7 arcsec, the performance of the telescope could be considered as seeing-limited if the telescope could reach its intrinsic quality of 0.30 arcsec FWHM.

Passive performance of an 8-m telescope. ► Figure 5-10a demonstrates how an active telescope with a primary mirror with a diameter of 8 m and a thickness of 175 mm would perform if it was operated passively. The data were obtained with continuous measurements of the coefficients of the low-order aberrations with the telescope following a star from the zenith to the horizon in a passive mode, that is, without applying any active optics corrections (Guisard et al. 2000). The figure shows the evolution of the coefficient of the y-component of the softest elastic mode $e_{2,1}$.

Obviously, the evolution can be split into a smooth evolution and fast random variations.

The systematic evolutions of the low-order coefficients were obtained by fitting, rather arbitrarily, sixth-order polynomials. The circles in ► Fig. 5-10b show the rms values of the systematic variations of the coefficients of the lowest 16 modes, plotted logarithmically because of the large differences between the modes. For random force errors, one should expect that the systematic



■ Fig. 5-9 Improvement of the image quality of the ESO 3.6-m telescope after the removal of low-order aberrations (Reproduced from Wilson 1982). Triangular coma is $\rho^3 \cos 3\phi$ and quadratic astigmatism is $\rho^4 \cos 4\phi$

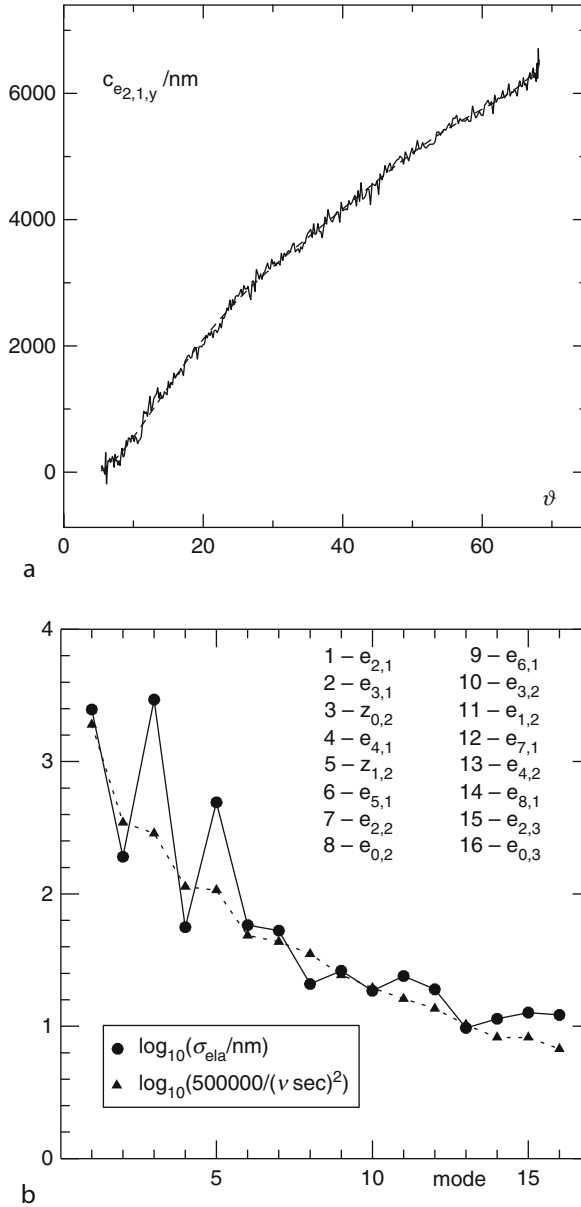


Fig. 5-10

(a) Coefficient of the y -component of the elastic mode $e_{2,1}$ as a function of the zenith angle ϑ . The *dashed line* is a best fit of a sixth-order polynomial. (b) Correlations between the scaled inverse stiffnesses of modes, expressed through their eigenfrequencies ν , and the measured rms values of the variation of the coefficients with the zenith angle

variations are inversely proportional to the stiffnesses of the modes. According to (► 5.2), these stiffnesses are proportional to the square of their eigenfrequencies ν listed in ► Table 5-2 in ► Sect. 5.1. They are represented in ► Fig. 5-10b by the triangles and are scaled to minimize the differences between the circles and the triangles for the lowest modes.

For most of the modes, the measured variations and the inverse stiffnesses are well correlated. Only the measured coefficients of $z_{0,2}$, which is Zernike defocus, and $z_{1,2}$, which is Zernike third-order coma, are larger than expected. The reason is that they are not generated by deformations of the primary mirror alone. The much stronger variations of these two coefficient are due to misalignments of the secondary mirror, which are caused by deformations of the telescope structure.

Nevertheless, for the other modes, the variations are inversely proportional to the stiffnesses and are probably generated by random forces. Otherwise, specific systematic force errors would have generated comparatively larger coefficients of specific lower modes. From the variation of 7,000 nm of the coefficient of the lowest elastic mode $e_{2,1}$ for a change of the zenith angle of 70° , one can estimate the rms of the random force errors. Statistically, the axial support may generate half of this error, that is $7,000 \text{ nm}/\sqrt{2} \approx 5,000 \text{ nm}$. This would require axial random force errors with an rms $\sigma_F \approx 30 \text{ N}$, which is equivalent to an rms of the variation of the nominal axial support forces of 2% and, therefore, close to what one would expect for the inaccuracies of the passive forces.

The conclusion of the measurement is that the wavefront errors that are generated during a passive operation of a telescope with a large, thin monolithic mirror exceed by far even the limits for seeing-limited operations.

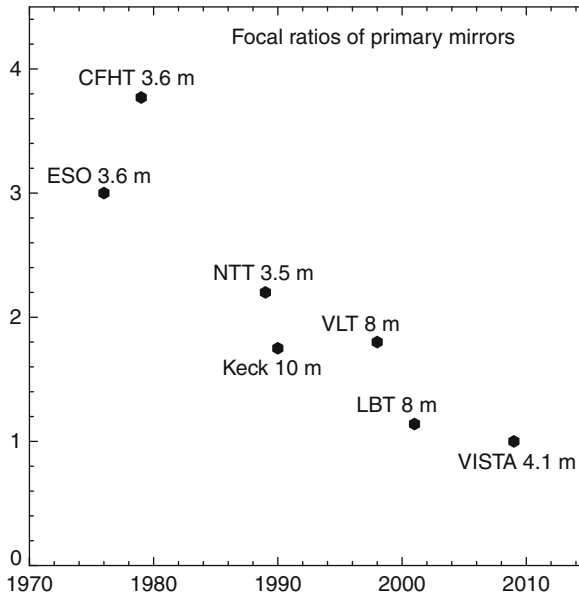
4 Active Optics in Telescopes with Meniscus Mirrors

4.1 Advantages of Active Optics

The limitations for the performance of large telescopes due to the deformation of the structure and application of incorrect forces can be overcome by a continuous control of the alignment and the shapes of the mirrors (Wilson et al. 1987). The most obvious disadvantage of thin mirrors, namely, that they change their shape under the influence of external forces, then turns into a major advantage. Low spatial frequency errors, which are difficult to control during the polishing of large mirrors, can be corrected by moderate forces during the operation of the telescope. As a consequence, certain amounts of low spatial frequency errors introduced by the polishing can be tolerated, and more emphasis can be put on minimizing the high spatial frequency errors.

Another important advantage is the capability to correct misalignments of the telescope mirrors. The sensitivity of the image quality to tilt and lateral misalignments of the secondary mirror increases strongly with the inverse of the focal ratio of the primary mirror and, therefore, with the inverse of the length of the telescope. The option to correct such misalignments during the operation of the telescope opens the door for the construction of much shorter and therefore lighter and more rigid telescopes. ► Figure 5-11 shows the evolution of the focal ratios of primary mirrors in the last 40 years. The CFHT and the ESO 3.6 m were typical passive telescopes of the 4-m class, whereas the NTT was the first active telescope.

With active optics, modifications of the optical configuration of the telescope also become feasible. One example is the switch between the telescope foci at the VLT. The optical design



■ Fig. 5-11

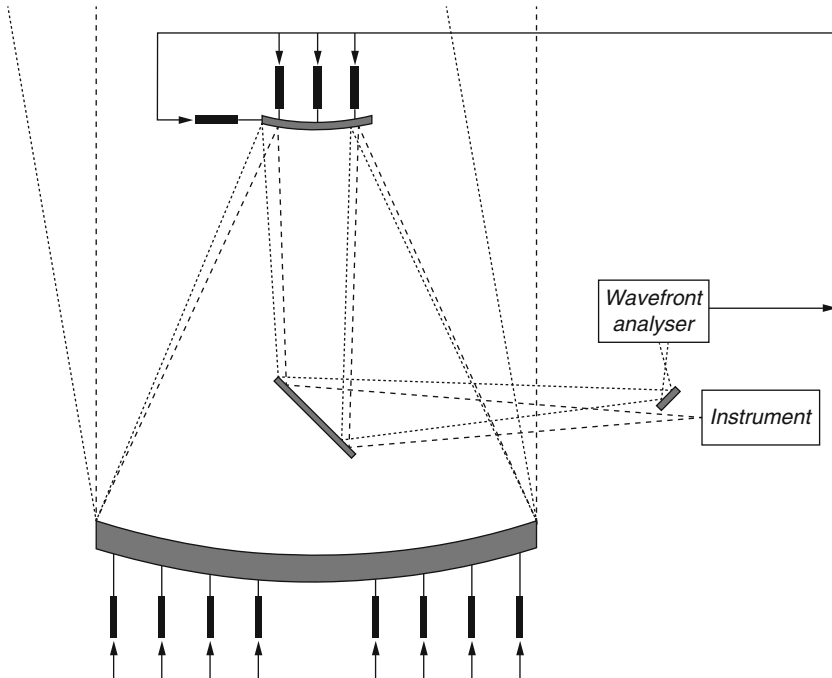
Evolution of the focal ratios of the primary mirrors in the last 40 years with the transition from passive to active telescopes

of the telescope is optimized for the Nasmyth focus and a switch to the Cassegrain focus with its different back focal distance requires a refocus with the secondary mirror. A consequence of this change of the first-order characteristics of the optics is a violation of the Ritchey-Chrétien condition, which generates strong aberrations even at the center of the field. However, by an appropriate deformation of the primary mirror, at least the dominant field-independent spherical aberration can be corrected. The major field aberration is then third-order coma with a linear dependence on the field angle. A correction of the coma would require a modification of the shape of the secondary mirror, which is not possible in the VLT.

Another example of a change of the optical configuration is the modification of the plate scale. In a two-mirror telescope with a flexible primary mirror, this can be done by introducing simultaneously two defocus errors that compensate each other. The first defocus is introduced by changing the radius of curvature of the primary mirror and the second one by an axial movement of the secondary mirror. Again, such a change of the first-order characteristics also generates spherical aberration and third-order coma, of which the first can be corrected by an additional modification of the aspheric shape of the primary mirror.

4.2 Control Strategy

Open-loop control. In principle, there are two possible types of control, namely, open- and closed-loop control. Both of them need signals to drive the correction devices (● Fig. 5-12). Ultimately, the signals are obtained by optical measurements of the wavefront aberrations at the focal stations of the telescopes, using reference stars as light sources.



■ Fig. 5-12
Principle of active optics corrections

Whereas open-loop corrections are based on calibration tables, which are obtained outside the observing hours, closed-loop operations measure the control signals during observations.

On the one hand, the advantage of open-loop control is that there is no need to operate the wavefront sensing during the observations. Furthermore, it can also be applied when sufficiently bright reference stars are not available.

On the other hand, the open-loop approach faces a few problems. First, often only a small number of measurements can be made, which leads to inaccurate calibration data if the measurements are strongly affected by noise. Second, effects like hysteresis cannot be calibrated and therefore limit the correction capabilities of the control system. Third, the components of the telescope system, in particular the passive supports and the force and position actuators, must be sufficiently stable and failure-free between the calibrations, possibly over periods of several weeks.

Closed-loop control. The alternative is to use closed-loop control, where the corrections are based on continuous measurements of the wavefront errors during the observations. A very important advantage of closed-loop control is the continuous check of the telescope quality, with the possibility of immediate corrective actions whenever a component fails. For example, if one support fails and applies incorrect forces, the generated errors can largely be compensated by other active supports. Such corrections could be based also on calibration tables. However, the setup of such tables would require reliable models for the prediction of the errors generated by failures of specific components and also sensors that detect the failures.

There are further advantages of closed-loop control compared with open-loop control. First, no observing time is lost to obtain calibration data. Second, the times over which components have to be stable can strongly be reduced. Third, as a by-product, issues such as the dependence of the image quality on temperature differences can be investigated statistically, based on possibly millions of measurements after a few years of operation (Cullum and Spyromilio 2000; Guisard et al. 2000). Such studies would be very difficult to undertake with only a few measurements since the measured parameters are often heavily affected by various types of noise.

Degree of control. Naturally, the amplitudes of the wavefront errors often increase rapidly with the size of the telescope. In addition, requirements on the resolution become more stringent in modern large telescopes, which are often designed to reach a diffraction-limited instead of a seeing-limited performance. The degree of control required to maintain the optical quality of the telescope then depends primarily on the size of the telescope, the efficiency with which it is stabilized by the passive components, and the desired image quality.

Multistage control. One important consequence of the increase of the telescope size is the decrease of the lowest eigenfrequencies of the full system. If the masses of the correcting devices are sufficiently large to excite the lowest eigenfrequencies, the bandwidth of the control will be limited by these frequencies. Yet, high bandwidth is required for a correction of fast disturbances as well as for an efficient correction of errors with large amplitudes.

One solution to this problem is to introduce a second stage of control, where the corrections are done by smaller elements. These can operate at higher frequencies without exciting vibrations in the elements with much higher masses. The small elements will nearly instantaneously correct all errors with high bandwidths. However, their strokes are usually not sufficient to correct the full amplitudes of the errors that evolve over longer periods of time. Therefore, well before they run out of their correction range, they have to offload the accumulated errors at lower frequencies to larger controlled elements. In addition, they can also be designed in such a way that the moving masses are balanced, thereby avoiding the generation of spurious vibrations.

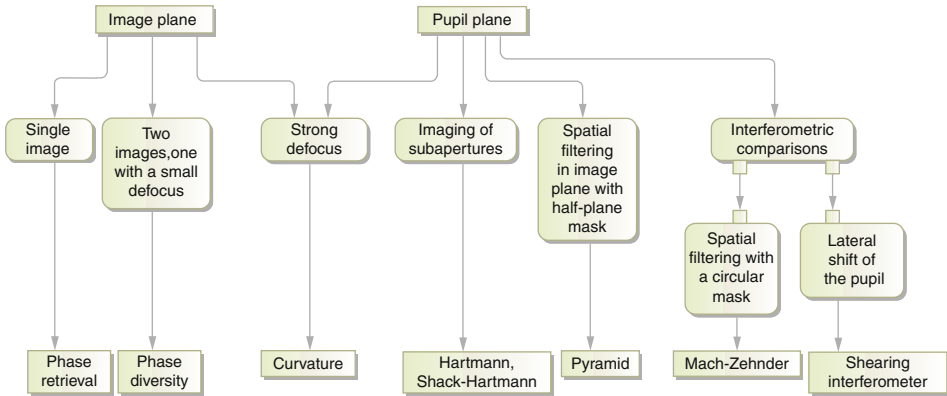
4.3 Optical Wavefront Sensors

4.3.1 Choice of the Type of Wavefront Sensor

Types of wavefront sensors. The wavefront sensors most commonly used in active optics can coarsely be classified according to the planes from which they retrieve the information on the wavefront aberrations: image planes or pupil planes (● Fig. 5-13). In the pupil planes, the phase information must be converted into detectable intensity information, which can be achieved by imaging of subapertures, by defocusing or by spatial filtering. One sensor, the curvature sensor, retrieves the information from planes somewhere between the image and the pupil planes.

The wavefront errors must be expressed as functions across the pupil. The pupil plane detectors have the advantage that the locations on the detector are directly related to the positions in the pupil and that the algorithms for the calculation of the wavefront aberrations therefore do not require iterations.

Image plane detectors. In principle, the wavefront aberrations can be deduced from the shape of the point images, a method called phase retrieval (Gerchberg and Saxton 1972; Gonsalves 1976). The major problem with this method is that different aberrations may have similar effects on the images.



■ Fig. 5-13

Classification of most common wavefront sensors used in active optics

To a large extent, the simultaneous analysis of a second image can remove the ambiguities. This method of disentangling the aberrations is known as phase diversity (Gonsalves 1982; Gonsalves and Childlaw 1979). The second image is obtained with an additional small amount of a well-known aberration, usually a defocus. However, close to the image plane, the effects are not linear, and the solution still requires an iterative process. Furthermore, the accuracy of the method is strongly affected by atmospheric effects.

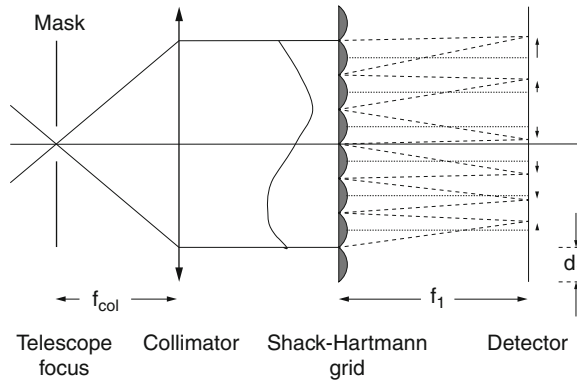
Curvature sensing. This method retrieves the wavefront errors either from strongly defocused images or strongly defocused pupil planes (Roddier 1988). The introduced amount of defocus must be sufficient to get out of the caustic near the image plane or sufficiently far away from the pupil plane to generate intensity variations. Curvature sensors measure the Laplacian, that is, the local curvatures of the wavefront error.

The information on the wavefront aberrations can be extracted directly for the locations in the defocused planes. Iteratively, these locations can be related to the positions in the pupil. Since aberrations like defocus and third-order astigmatism have constant curvature, their coefficients cannot be determined from information inside the defocused image. Instead, their coefficients are derived from the size and the shape of the edge of the image.

This type of wavefront sensing is of interest in particular in survey telescopes, where the whole field is often covered by the detector. By displacing a few of the chips of the detector along the optical axis up or down, the images of the stars in the corresponding fields are defocused and can be used for this type of wavefront analysis.

Shack-Hartmann sensor. The most common pupil sensor is the Shack-Hartmann wavefront analyzer (Platt and Shack 1971; Shack and Platt 1971; Wilson 1999) (● Fig. 5-14), in which a lenslet array is placed in a reimaged exit pupil. Each lenslet samples a subaperture in the pupil and creates an image of the star on the detector. In the focal plane of the telescope, a mask with a diameter of only a few arcseconds usually ensures that the spot pattern is generated by the light of a single star.

The differential movements of the spots in the focal plane of the lenslets are proportional to the average slopes of the wavefront across the corresponding subaperture in the pupil. It would then be natural to fit derivatives of functions. If only the coefficients of Zernike polynomials were required, one could fit the derivatives of Zernike-type polynomials, whose derivatives are



■ Fig. 5-14

Principle of the Shack-Hartmann wavefront sensor

orthogonal over the unit circle (Lukosz 1963). However, there are no orthogonal sets of derivatives for other functions like the elastic or segment alignment modes. A better solution is then to fit, after a reconstruction of the wavefront from the local slopes, the functions themselves.

The lenslet array may be irregular, and additional aberrations may be generated by the optics of the sensor itself. The common solution to single these errors is to compare the spot pattern generated by the reference star in the sky and the full optical train with a pattern generated by a perfect wavefront and the optics of the sensor. The perfect wavefront can be produced by an artificial point source at a telescope focus in front of the sensor.

The overall tip and tilt of the wavefront is given by the average positions of all spots. These also depend on the position of the mirror, which picks up the light of the reference star and whose position is often not known with high accuracy. Therefore, tip and tilt cannot be measured with high absolute accuracy and are not usually included in the set of measured modes. Furthermore, the focus in the instrument may not be equivalent to the focus inside the wavefront analyzer. However, once this offset between the two foci has been calibrated, the wavefront analyzer can also measure the focus errors seen by the instruments.

The number of subapertures should be at least as large as the number of controlled modes. The choice of the size of the detector and the diameter and the focal ratio of the lenslets is a compromise between the effects of several parameters (Noethe 2002). For example, for faint stars, the size of the spots should ideally be approximately twice as large as one pixel. In addition, a spot should not interfere with neighboring spots. Both features require small spot sizes and therefore small focal ratios. On the other hand, larger focal ratios increase the sensitivity to slope errors.

In general, a major advantage of the Shack-Hartmann method is its simplicity in terms of analysis. Furthermore, nearly all the light is used and concentrated on a relatively small number of pixels, which improves the signal-to-noise ratio.

Pyramid sensor. Another sensor, which measures slopes in the pupil, is the pyramid wavefront sensor (Raggazoni 1996), which is a two-dimensional implementation of a linearized knife-edge test. In a traditional knife-edge test, a sharp blade covers a fixed half plane in the focal plane. In the geometrical optics regime, such an arrangement only gives information on whether a ray passes through the uncovered half plane or not. However, by moving the edge up and down, the fraction of the time when the ray passes through the unobstructed half plane depends smoothly

on the location where the ray intersects the focal plane. Since this location depends linearly on the slope of the ray in the pupil plane, the intensity of the corresponding pixel in a reimaged pupil plane on the detector is proportional to the slope.

The use of a prism, with the edge taking over the role of the knife edge, offers the advantage of sampling the two half planes simultaneously and thereby using all of the light. Furthermore, replacing the prism by a glass pyramid yields information on the slopes of the wavefront in two orthogonal directions. One advantage of the pyramid sensor is that its sensitivity can be adjusted to the actual image quality by adapting the stroke of the movements to the size of the image in the focal plane.

Interferometric sensors. Interferometric methods also usually work in pupil planes. The wavefront that will be compared interferometrically with the wavefront in the exit pupil of the telescope can be generated by spatial filtering with a circular mask in the image plane, as in a Mach-Zehnder interferometer. An alternative is a lateral shift of the pupil as in the shearing interferometer, which then measures slope errors.

Accuracy. The accuracy of a wavefront sensor depends to a great extent on nuisance parameters related to the detector. For a CCD, in particular, these are the readout noise and the non-uniformity of the sensitivity of individual pixels. A common practical assumption is that the rms σ_{pixel} of the centroiding errors is of the order of 5% of the pixel size, caused to a large extent by the pixelization itself.

Particularly for faint stars, a major source for centroiding errors is the photon shot noise. With only a small number of photons available, the best centroiding precision is achieved if the light is, in one dimension, concentrated on two pixels. In pixel units, the rms of the spot size is then approximately equal to 1. Now, for a single photon, the rms σ_{photon} of the centroiding error due to photon shot noise is equal to the rms of the spot size. Therefore, a reduction of σ_{photon} to a value similar to σ_{pixel} , that is, to approximately 0.05, requires about 400 photons.

For subapertures with diameters of half a meter, typical transmissions of the atmosphere and the optics, a quantum efficiency of the detector of more than 50%, and a bandwidth of the light of a few hundred nm, 400 photons would be delivered by stars of magnitude of the order of 16. The feasibility to use such dim stars nearly guarantees a full sky coverage even close to the galactic pole.

Atmospheric noise. If only those errors are measured and corrected that are due to slowly varying error sources like the change of the altitude angle, the disturbing effects of the free atmosphere can be reduced by longer integration times. The chosen integration time and therefore the minimum time between corrections also depend on the rate of change of the wavefront errors introduced by the various error sources. In pure active optics, typical integration times are of the order of 1 min.

4.3.2 Indistinguishable Error Sources

Ideally, any error generated by a specific optical element of the telescope would be corrected by the same element. One necessary condition for this is that the wavefront sensing is capable of attributing the errors unambiguously to the specific elements. For a single field location, the contributions of different elements to any of the modes simply add up. With only one wavefront sensor using a star in only one field position, the error sources can, therefore, not be disentangled.

However, most errors generate aberrations in the field with specific dependencies on the field locations. The amplitudes of these field aberrations depend on the distance of an optical element from a pupil. Therefore, as long as two elements are not optically conjugated, it may be possible to disentangle the error sources. However, this requires that the coefficients of a sufficient number of modes can be measured at different field positions and that any two error sources always generate different field aberrations.

A major problem would still be the measurement of distortions of the field, which require the precise knowledge of the position of the star in the sky and the wavefront sensor in the focal plane of the telescope. Often, the latter is not known with sufficient accuracy, and, therefore, information on the distortion of the field cannot usually be obtained.

Misalignments. An unambiguous correction of a misalignment of N mirrors requires $2(N - 1)$ modes to be measured with more than one wavefront sensor in the field (Piatrou and Chanan 2010; Schechter and Sobel 2010). Apart from the first-order aberrations tip, tilt, and defocus, misalignments of the mirrors generate predominantly third-order aberrations: spherical aberration for movements along the optical axis and third-order coma and third-order astigmatism for lateral and rotational out-of-plane movements. The only remaining third-order effects, namely, the field distortions, are difficult to measure. Also the coefficients of the fifth- and higher-order optical aberrations are usually too small to be measurable with sufficient accuracy.

The simplest and also most important example of how to use a certain set of modes and a minimum number of wavefront sensors is the alignment of a two-mirror Ritchey-Chrétien telescope. A perfect Ritchey-Chrétien telescope is, in addition to spherical aberration, free of third-order coma in the field. This is still the case after a rotation of its secondary mirror around the coma-free point, which is located approximately halfway between the center of curvature and the apex of the secondary mirror. Freedom of coma therefore only guarantees that the axes of the two mirrors intersect in the coma-free point, but not that the telescope is fully aligned.

The full alignment therefore requires the measurement of another mode, whose coefficients in the field depend on the angle between the two axes. The one with the largest coefficients is third-order astigmatism, for which the misalignment generates a linear field dependence. A Ritchey-Chrétien telescope can therefore be perfectly aligned by measuring two modes, namely, third-order coma and third-order astigmatism, with at least two wavefront sensors (McLeod 1996; Noethe and Guisard 2000).

Figure errors. In addition to the ones generated by misalignments, wavefront aberrations are also generated by figure errors of the mirrors. If a mirror is not located in a pupil, the figure errors will also give rise to additional field aberrations.

Such additional field aberrations due to figure errors in the form of third-order astigmatism are all distortions. Since these are difficult to measure, the contributions of the third-order astigmatism of the individual mirrors to the total amount of this mode cannot usually be disentangled. The same applies to the contributions to third-order coma from misalignments of the mirrors on the one hand and figure errors of the mirrors in the form of this mode on the other hand.

As a consequence, in both examples, the same wavefront aberrations can be corrected by different combinations of correcting elements. However, if such ambiguities cannot be detected by the wavefront sensors, they will, most likely, not have any deleterious effects on the image quality. The only remaining error that may have an impact on the image quality will then be a distortion of the field.

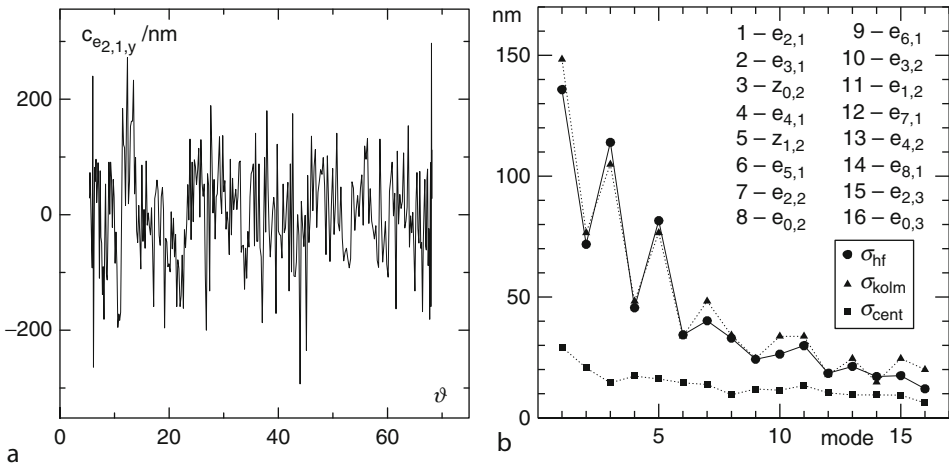
4.3.3 Disturbances by the Atmosphere

Consider again the passive evolution of the coefficient of the lowest elastic mode $e_{2,1}$ of symmetry 2 with the altitude angle (● Fig. 5-10a). The residual after the subtraction of the systematic evolution represents the high temporal frequency variations (● Fig. 5-15a). The circles in ● Fig. 5-15b show the rms values σ_{hf} of the high-frequency residuals for all the modes controlled in the VLT.

For comparison, the squares represent the coefficients σ_{cent} that one would expect from random centroiding errors with rms values of as much as 30% of the pixel size. Obviously, not only the amplitude, but also the relative values of the coefficients expected from centroiding errors are very different from the corresponding measured figures σ_{hf} . Therefore, the accuracy of the measurements of the low-order coefficients generated by the telescope optics is not limited by the accuracy of the wavefront analysis.

Instead, a good correlation can be established with the variation of the coefficients expected from the Kolmogoroff model of atmospheric turbulence (Fried 1965; Noll 1976). For a given diameter of the telescope, these variations of the coefficients depend only on the Fried parameter r_0 , which has a value of approximately $r_0 = 130$ mm for median seeing conditions of 0.7 arcsec.

The Fried parameter takes variations of all frequencies into account, but for longer integration times, the high-frequency contributions are averaged out. One may then define a fictitious Fried parameter \bar{r}_0 for the integration times of 30 s used for the measurement. With a value of $\bar{r}_0 \approx 700$ mm, the expected variations σ_{kolm} of the coefficients, represented by the triangles, are in excellent agreement with the measured variations σ_{hf} . This suggests that, at least for the lowest modes, the noise is predominantly generated by atmospheric turbulence.



■ Fig. 5-15

(a) Residual of the coefficient of $e_{2,1}$ as a function of the zenith angle ϑ after removing the smooth curve shown in ● Fig. 5-10a. (b) Correlations between expected rms values of the variations of the coefficients due to centroiding errors (σ_{cent}) or Kolmogoroff turbulence (σ_{kolm}) and the rms (σ_{hf}) of the measured high-frequency nonelastic variations

4.4 Correction of Errors

4.4.1 Correction of Misalignments

Misalignments can be corrected by appropriate rigid-body movements of the optical elements. In two-mirror telescopes, the major device for such corrections is usually the secondary mirror. A connection of the structure holding the secondary mirror cell to the telescope structure by a hexapod facilitates rigid-body movements in all degrees of freedom.

The ultimate reference for the alignment is the plane defined by the instrument adapter. Ideally, the telescope axis should intersect this plane at the center of the adapter and should be perpendicular to it. A full alignment of a two-mirror telescope to the instrument would then also require a control of the rigid-body positions of the primary mirror. Whether this is possible depends on the design of the fixed points. With a hydraulic support system, the position of the primary mirror can rather easily be controlled in five degrees of freedom. On the other hand, for a segmented mirror, a simultaneous lateral movement of all segments is very difficult to realize.

4.4.2 Correction of Mirror Shapes

Comparison of the correction of elastic and Zernike modes. Since most of the deformations in the mirrors will be generated by incorrect support forces, the deformations are expected to be in the form of elastic modes. For the correction of mirror figures, it is, therefore, more natural and efficient to generate deformations in the form of the elastic modes of the mirrors rather than in the form of Zernike polynomials. Both the required forces and the residual errors will be much smaller if elastic modes are used for the corrections.

A good example for a higher accuracy obtained with smaller forces is the comparison of the forces that are required to generate the lowest elastic mode $e_{2,1}$ and the corresponding Zernike polynomial $z_{2,1}$, which represents third-order astigmatism. The rms of the relative difference between these two modes is only approximately 5%. In the VLT mirror, 1,000 nm of the elastic mode can be generated with a force pattern with a maximum force of $F_{\max} = 1.68$ N on one of the actuators and with a residual error of 0.03 nm rms. Despite applying a much higher maximum force of $F_{\max} = 13.6$ N, the Zernike mode can only be generated with a residual error of 2.4 nm rms.

An expansion of the Zernike mode in elastic modes of rotational symmetry two and a correction of only a few of the elastic modes would also reduce the maximum forces. For example, approximating the Zernike mode by the two lowest modes only, the maximum force would be reduced to 4.2 N. However, the residual error would increase to 12 nm.

Choice of set of corrected modes. The choice of active modes, that is, those to be corrected, depends on the characteristics of the optical elements of the telescope. Misalignments of optical elements will generate considerable amounts of Zernike defocus and third-order coma, which should therefore be included in the set of active modes.

For a telescope with a large monolithic primary mirror, the natural choice for the other active modes are the elastic modes of this mirror. However, two of them, the lowest elastic modes $e_{0,1}$ and $e_{1,1}$ of the rotational symmetries 0 and 1, will not be included in the set of active modes since they are very similar to the two corresponding Zernike modes defocus and third-order coma, respectively.

How many of the other elastic modes should be included depends on the stiffness of the mirror, the available force range, and, possibly, on the noise in the wavefront measurements. The forces required for the correction of the modes increase strongly with their spatial frequencies. Because of the limited range of the correction forces, this prohibits the correction of higher-order modes unless their coefficients decrease so rapidly that the correction forces are sufficiently small. Usually, this is not the case in wavefronts affected by various types of noise, in particular atmospheric noise.

One way to define the set of active modes starts with the assumption that most of the errors in the mirror are generated by force errors. From experience, the force errors are random with an rms of approximately 2% of the nominal force at zenith. Such random forces generate deflections in all modes, which, however, decrease strongly with the order of the mode. Therefore, corrections are applied only to those modes that are generated with coefficients larger than a small fraction of the diffraction limit, say 15 nm for visible wavelengths. The coefficients can conveniently be calculated by the methods described in [Sect. 2.3.5](#).

Potential use of lateral supports for corrections. The easiest way to control the shape of a mirror is to modify the axial support forces. In addition, modifications of the lateral support forces could be used. However, since these supports can only generate and therefore correct one specific deformation for each rotational symmetry, the control of mirror shapes is usually only done with the axial supports.

Scaling laws for active mirrors. The first scaling of interest is the one for the forces required to generate and therefore correct the aberrations that are present in the initially uncorrected, that is, passive telescope. Using the scaling for the number n_{sag} of supports that are required to fulfill the specification for the sags between the supports ([Sect. 3.4.4](#)) and the scaling h^3/D^2 to generate a given deformation with a single force, the active forces $F_{a,c}$ required to generate global wavefront errors with identical rms values scale very strongly with $(h/D)^4$. However, the active forces $F_{a,p}$ required to correct the expected wavefront errors due to incorrect passive forces, which scale with $(D^2/h)^{3/2}$, scale more weakly with $h^{5/2}/D$.

The second scaling of interest is the one for the tolerable errors in the active forces, which will finally limit the quality of the mirror in terms of low spatial frequency aberrations. Assuming that the errors in the n_{sag} active forces are random and proportional to the active forces $F_{a,p}$ and using that the effects of a single force on the wavefront error scale with D^2/h^3 , the generated wavefront errors scale with $(h^{5/2}/D)\sqrt{D^2/h}(D^2/h^3) = D^2/h$. For a given wavefront error, the tolerances for the active force errors therefore scale with $F_{a,\text{err}} \propto h/D^2$ and, therefore, decrease strongly with the diameter of the mirror than the absolute tolerances.

However, the mechanical limitations for the accuracy of the force setting are not absolute values, but are given by fractions of the total force. Therefore, if the setting of the active forces is independent of the passive forces, the scaling of the relative tolerances of the active force errors is obtained by dividing h/D^2 by the maximum $F_{a,p}$ of the active forces. The relative tolerances then scale with $(h/D^2)/F_{a,p} \propto (h/D^2)/(h^{5/2}/D) = 1/(Dh^{3/2})$ and, therefore, more weakly with the diameter of the mirror than the absolute tolerances.

Often, the maximum active forces are higher than $F_{a,p}$, since the active forces are also used for the correction of polishing errors and for modifications of mirror shapes that are required for a change of the optical configuration ([Sect. 5.2.1](#)). Nevertheless, the active forces are usually still much smaller than the passive ones and should be decoupled from the passive forces.

Fixed points. In axial supports with astatic levers, three support points are usually chosen as fixed points. They are preferably located at angles of 120° . Of course, the forces at these points cannot be modified by actuators. However, when correction forces with a certain rotational symmetry

are applied at all other support points, the three conditions of the equilibria of the forces and the two moments guarantee that the reaction forces on the three fixed points automatically assume the values requested for the correction.

Number of supports required for the correction. The required total number of supports depends on the set of modes that have to be corrected and on the purity with which the modes have to be generated. Figure 5-16a, b show the residual rms σ_{resid} that can be obtained if a mode i of the rotational symmetry 2 and an rms of 1 nm is corrected with n rings. At least partially, n rings can correct modes up to the order $n - 1$.

For a given order i of the mode, ranging from 1 to 5, Figure 5-16a shows σ_{resid} as a function of the number n of rings. The slopes in the log-log plot are approximately -4 . For a given number n of rings, ranging from 1 to 6, Figure 5-16b shows σ_{resid} as a function of the orders i . The slopes in the log-log plot are approximately $+6$. Therefore, the residual error σ_{resid} scales approximately with i^6/n^4 .

Because of the existence of three fixed points, the generation of all modes with rotational symmetries other than multiples of three will also generate a tilt of the mirror. During observations, this tilt will be quickly corrected by an autoguider working with correction rates of at least a few Hertz.

Furthermore, a limited number n_0 of supports on a ring will cause crosstalk between modes. If the system generates a mode with a rotational symmetry $m < n_0$, it will also generate deformations with the rotational symmetry $n_0 - m$. Generating a more rigid mode with a rotational symmetry m close to n_0 may then generate considerable crosstalk into a much softer mode with the much lower rotational symmetry $n_0 - m$. However, in the next active optics cycle, this crosstalk will be eliminated by a correction of the lower mode, and the generated crosstalk into the higher mode will be insignificant (Noethe 2002).

Corrections with position-based supports. In principle, a meniscus mirror can also be actively supported by position actuators. However, with such a position-based support the mirror is

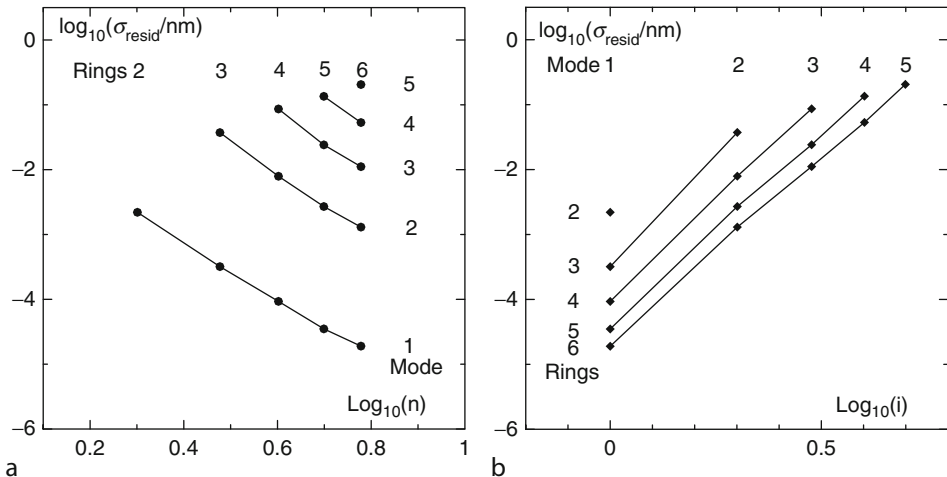


Fig. 5-16

Purity of generated modes of rotational symmetry two. (a) Mode given, dependence on the number of rings. (b) Number of rings given, dependence on the order of the mode

strongly coupled to the mirror cell. Because of potentially large deflections of the mirror cell, corrections have to be applied more frequently than in support systems with force actuators, where the mirror can largely be decoupled from the mirror cell.

Large mirror cells deform in the lowest-order modes by approximately 1 mm peak-to-valley between horizontal and vertical positions. With a position-based support, the maximum rate of change of the coefficient of the lowest-order mode will then be of the order of 50 nm/s, which would require correction rates of the order of 1 Hz. During time intervals of 1 min, the corrections could be done at this rate in open loop based on calibration tables. Measurements of the wavefront errors with integration times of 1 min could then perform closed-loop corrections of possible open-loop errors and update the calibrations tables accordingly.

One case, where a position-based support is unavoidable, is the support of a segmented mirror. Each individual segment is usually supported by three whiffle trees. The bases of these whiffle trees define the rigid-body position of the mirror with respect to the mirror cell. Therefore, seen as one unit, a segmented mirror has a position-based support. Consequently, corrections of the errors caused by the large deformations of the mirror support structure down to the nanometer level require correction rates of a few Hertz.

4.4.3 Correction of Deflections of the Structure

Most of the tubes of telescopes up to the 8-m class are supported by bearings close to the altitude axis. Larger telescopes like radio telescopes often use rocking-chair designs with large rails closer to the center of the tubes, which serves to reduce the deflections of the tube and in particular the primary mirror cell. It would also be possible to extend the concept of active optics to the structures, that is, to introduce active devices that control the shape of the structures.

4.5 Historical Development of Active Optics

First suggestions of a systematic process to improve the image quality, in particular by modifying the forces under the primary mirror, were made by Couder (1931) and Maksutov (1954). Much later, the control of the shape of a thin, deformable primary mirror was suggested for an orbiting astronomical observatory (Creedon and Lindgren 1970), using the elastic modes of the mirror. With a chosen number n_{supp} of supports, all $n_{\text{supp}} - 3$ deformation modes that could be generated by these supports were controlled. In addition, to reduce the deformations in the weakest modes above the set of controlled modes, the supports were located as closely as possible to the nodes of these weakest modes. An experiment with a mirror with a diameter of 790 mm verified the feasibility of this approach.

Independently, a second control scheme was derived from the results of the measurements at the ESO 3.6-m telescope described in ► Sect. 3.5. The major difference to the other scheme is that the number of controlled modes is much smaller than the number of supports (Wilson 1982). Although this approach does not yield the best possible correction, it reduces the forces by correcting only those modes, in which the mirror can easily be deformed. Usually, these modes are also the dominant modes in the expansion of the wavefront error. The distribution of the support points is then primarily driven by the geometry of the mirror and the requirements on the sag between the supports.

5 Active Telescopes

5.1 Telescopes of the Four-Meter Class

Dimensions of the Primary Mirror. The first telescope that was right from the start designed and operated as an active telescope was the New Technology Telescope of the European Southern Observatory (ESO) (Wilson et al. 1991) with a diameter of the primary mirror of 3.58 m. At that time, active optics was a new, not yet proven technology. Therefore, in addition to the active mode, the telescope was requested to be capable of working also in a passive mode, albeit with a reduced performance.

Consequently, the thickness of 241 mm for the primary mirror (► Fig. 5-17) was a compromise between a typical thickness of 500 mm of a passive mirror and a possible thickness of 100 mm of a fully active mirror. Nevertheless, the chosen thickness reduced the weight of the NTT primary mirror to half of that of a conventional mirror with the same diameter. After the operation of the NTT had proven the feasibility of active optics, it was applied in all new telescopes with monolithic mirrors of similar or larger size.

Optical performance specifications. The NTT had to fulfill performance specifications both for the passive and the active mode. The performance was defined in terms of the diameter d_{80} of the circle containing 80% of the points in the geometrical point spread function generated by the telescope errors alone. The specifications were $d_{80} = 0.4$ arcsec for the passive and $d_{80} = 0.15$ arcsec for the active mode, which is equivalent to image sizes of approximately 0.3 and 0.1 arcsec FWHM, respectively.

Passive components. Contrary to most of the classical predecessors in the 3–4-m class and following the evolution of telescope mounts in that period, the NTT has an alt-azimuth mount. In particular, this mount simplified the design of the lateral support of the mirrors since gravity is acting in only one direction relative to the mirror.



► Fig. 5-17
NTT mirror in the optical workshop

■ Table 5-2

Eigenfrequencies, given in Hz, of the lowest elastic modes of the NTT and the VLT

Symmetry	2	3	0	4	1	5	2	0	6
Order	1	1	1	1	1	1	2	2	1
NTT	115	273	192	479	434	732	737	852	1,034
VLT	16	38	42	66	68	102	107	119	143
Symmetry	3	1	7	4	8	2	0	5	3
Order	2	2	1	2	1	3	3	2	3
NTT	1,131	1,229	1,383	1,577	1,779	1,749	2,050	2,077	2,366
VLT	160	176	192	221	246	246	272	289	331

Like conventional passive telescopes, the NTT is equipped with all the passive devices that stabilize the image quality in the presence of potential error sources. For example, the structures holding the cells of the primary and secondary mirrors are attached to the center piece via Serrurier trusses.

Following the rotational symmetry of the primary mirror, the support points were placed on concentric rings. Due to the sag between the supports points, a three-ring support would have generated high spatial frequency slope errors of the order of 0.05 arcsec, which would have been too close to the overall specification of 0.1 arcsec FWHM for the active mode. Therefore, a support was chosen with 9, 15, 24, and 30 supports on 4 rings, which generated slope errors of 0.025 arcsec FWHM. 75 out of the 78 supports are astatic levers. The remaining three supports, which are located on the third ring at relative angles of 120° , are the fixed points.

The combination of the thickness and the curvature is such that the plane perpendicular to the optical axis containing the center of gravity intersects the rim. Therefore, the lateral support points could be located on the rim in this plane without the need for axial force components. The equidistant forces are all identical and point in the direction of the projection of the gravity vector onto the plane of the lateral support points. More details are given in [Sect. 3.4.4](#) and in [Fig. 5-7](#).

Equipment for active operation. In many respects, the NTT represents a straightforward evolution from older types of telescopes. The additional equipment required for an active operation consists of a wavefront sensor, a motorized position control of the secondary mirror, and a motorized control of the positions of the counterweights on the astatic lever supports under the primary mirror ([Fig. 5-12](#)). The active forces are therefore coupled to the passive support forces.

The wavefront sensor is of the Shack-Hartmann type, with a 25 by 25 lenslet array, that is, each lenslet collects the light of a 140 mm by 140 mm subaperture. The sensor is installed in the guide probe and can measure the wavefront errors continuously, that is, also during observations. It uses guide stars in the outer annular section of the full 30 arcmin field.

The secondary mirror can be moved along the optical axis to correct the focus of the telescope. Furthermore, it can be rotated around the center of curvature to correct the field-independent third-order coma without affecting the pointing. As described in [Sect. 4.3.2](#), such a coma correction does not fully align the telescope optics. It only guarantees that the axes of the primary and secondary mirror intersect in the coma-free point. A full alignment requires both the measurement of the coefficient of third-order astigmatism in at least two field positions and the capability to rotate the secondary mirror around the coma-free point.

With only one wavefront analyzer available, such measurements at different field positions of the coefficient of third-order astigmatism could only be done sequentially. This procedure

proved to be very tedious and time-consuming, in particular due to noise introduced by the atmosphere. Furthermore, the correction, that is a required rotation of the secondary mirror around the coma-free point, could only be performed by installing shims to shift the cell of the secondary mirror with respect to the telescope structure. Therefore, this alignment has so far only been performed once (Gitton and Noethe 1998).

To enable changes of the support forces for the correction of the shape of the primary mirror, only a minor modification had to be done to the axial supports used in passive telescopes of this size. Instead of using the usual astatic levers with fixed counterweights, the corresponding active astatic levers were equipped with motorized movable counterweights. They can adjust the support forces within a range of $\pm 30\%$ of the nominal support forces.

Since these nominal forces are proportional to the cosine of the zenith angle, the range for the correction forces is strongly reduced for large zenith angles. Therefore, in addition, the active supports were equipped with manually adjustable springs, which supply correction forces that are independent of the inclination of the telescope tube.

Corrected modes. The number of modes that have to be corrected can be calculated with the method introduced in [Sect. 4.4.2](#). The stiffest elastic mode that random force errors with an rms of 2% of the nominal force of 750 N at zenith can generate with a coefficient larger than 15 nm is the mode $e_{3,1}$, the first mode of rotational symmetry 3. Therefore, one would need to correct only the two softest modes $e_{2,1}$ and $e_{3,1}$. However, the next stiffest mode $e_{4,1}$ has also been included in the set of active modes.

Four rings are sufficient to correct the modes $e_{2,1}$, $e_{3,1}$ and $e_{4,1}$ with accuracies of at least 98%. According to [Fig. 5-16a](#), the lowest mode $e_{2,1}$ of rotational symmetry 2 can even be corrected with a residual relative error of 10^{-4} .

The lowest elastic modes of the rotational symmetries 0 and 1, which are similar to defocus and third-order coma, are not corrected by deforming the primary mirror. Instead, they can be corrected with sufficient accuracy by rigid-body movements of the secondary mirror.

The limit for the rms of the wavefront error introduced by random incorrect settings of the active forces was very stringently defined as 30 nm rms. However, this performance can be achieved with a rather relaxed accuracy of the force setting of the order of 3 N rms. The same value of 30 nm was also specified for the wavefront error introduced by imperfect corrections of defocus and third-order coma, using rigid-body movements of the secondary mirror along the optical axis and rotations around its center of curvature, respectively. The accuracies for the differential axial movements have to be of the order of 4 μm and for the differential rotations of the order of 6 arcsec. The latter involves differential lateral movements of the secondary mirror of approximately 120 μm and is therefore rather easy to achieve. The most difficult mode to control in the NTT is therefore the defocus.

Correction of polishing errors. Like the Hubble Space Telescope, the primary mirror of the NTT initially suffered from a figuring error, which had been introduced by an incorrectly assembled null system. The wavefront error was predominantly third-order spherical aberration ρ^4 with a coefficient of the order of 3.5 μm , which is equivalent to a FWHM in the point spread function of 0.5 arcsec. In the passive mode, the telescope image quality would therefore have exceeded its specification of 0.3 arcsec FWHM. Usually, such an error would have required a refiguring of the primary mirror or, possibly, of the secondary mirror. However, an inclination-independent correction of the first elastic mode of rotational symmetry 0 by the springs included in each of the active supports and a simultaneous correction of defocus could nearly perfectly correct this error.

The axial supports are pure push supports. The maximum negative active correction force that a support can apply at a certain zenith angle is equal to its nominal force at this zenith angle. Otherwise, if a negative correction force at a certain support is larger than the nominal force at this support, the mirror will be lifted off this support. Since the nominal forces are proportional to the cosine of the zenith angle, a need for negative correction forces will limit the zenith angle, up to which corrections can be performed. In the NTT, the strong active forces required for the correction of the polishing error restrict the maximum zenith angle to 75° .

5.2 Telescopes of the Eight-Meter Class

5.2.1 Design of the VLT Active Optics System

Dimensions of the primary mirror. The next step in the evolution of telescopes with large thin monolithic mirrors, both with respect to the size of the primary mirrors as well as the role of active optics, was the ESO Very Large Telescope (VLT), an array of four identical telescopes. The masses of the primary mirrors with diameters of 8 m and thicknesses of 175 mm are approximately 23 t (see Fig. 5-18). In terms of the response of wavefront errors to passive force errors and assuming a design of the support system with identical sags between the support points, the VLT is approximately 30 times more flexible than the NTT (see Sect. 3.4.4), which excluded a passive operation of the telescope right from the start.

Optical performance specifications. The VLT was designed as a telescope with a seeing-limited performance. Expressed in terms of the central intensity ratio CIR (see Sect. 2.2), the global specification was $\text{CIR} = 0.8$ for seeing conditions of 0.4 arcsec (Dierickx 1994). The two largest contributions were allocated to the surface errors after active optics corrections with a CIR of 0.920 and the control errors due to guiding and active optics with a CIR of 0.944.

The surface errors were split into the deformations generated by the sags between the supports and the high spatial frequency residuals of the generated wavefront errors after the active



Fig. 5-18

Back surface of a VLT primary mirror with the 150 tripod supports

optics corrections. For the primary mirror the specification for the sag between the supports was $CIR = 0.992$, which corresponds to a FWHM of the point spread function of 0.02 arcsec or to an rms of the wavefront print-through of 16 nm. To reach this specification for the high-frequency residuals, the manufacturer could use maximum active forces of 100 N under the primary mirror to remove the low spatial frequency errors, which are unavoidable with such a flexible mirror.

The specification for the active optics errors was $CIR = 0.979$, which was more or less evenly split into contributions from the wavefront sensing, the coma and defocus corrections by the secondary mirror and errors due to the setting of the active forces under the primary mirror. For each of these contributions, the related CIR of 0.995 corresponds to an rms of the dominant low spatial frequency modes of approximately 100 nm.

In terms of wavefront errors, the specification for the high spatial frequency errors seems to be more stringent than the one for the low spatial frequency errors. However, the central intensity ratio is a metric that is designed for the geometrical optics regime, in which the image quality depends primarily on slope errors. For a given rms of the wavefront errors, these slope errors are much larger for the high spatial frequency wavefront errors than for the low spatial frequency ones.

Passive components. As in the NTT, the structures holding the cells of the primary and secondary mirrors are connected to the centerpiece via Serrurier trusses. The design of the passive axial support system was driven by the specification of 16 nm for the rms of the sag between the support points. Four hundred simple support points would have been required to fulfill this specification, which was considered too complicated and, as shown below, was also not necessary for the correction of the lowest elastic modes.

Instead, the number of active supports was restricted to 150. Each of these supports distributes the force to three points via a tripod (Schneermann et al. 1990) (● Fig. 5-18). The supports are arranged on six rings, with 9, 15, 21, 27, 36, and 42 supports, respectively. Mechanical astatic levers would have added an appreciable amount of weight to the mirror cell. Instead, hydraulic systems were chosen for the passive axial and also the 64 lateral supports of the VLT primary mirrors.

Equipment for active operation. As is the case with the NTT, the VLT is equipped with only one Shack-Hartmann wavefront sensor with a sampling of approximately 20 by 20 subapertures. Unlike the NTT, electromechanical actuators apply the active forces via a spring in series with the passive forces. Consequently, the active forces are independent of the passive ones and, therefore, also independent of the inclination of the telescope tube.

For the control of five degrees of freedom for the rigid-body position of the primary mirror, the volumes in each of the three sectors of the passive hydraulic support as well as in the two sectors of the lateral hydraulic support can be modified. Corrections are done in closed loop whenever the mirror is laterally more than 100 μm out of position with respect to its cell.

The rigid-body position of the secondary mirror can be controlled in 5 degrees of freedom. Therefore, in principle, the axis of the secondary mirror can be fully aligned with the axis of the primary mirror. During operation, this would require the use of two wavefront sensors, whereas only one is available. However, a full alignment was done during the commissioning, based on simultaneous measurements with the permanently installed wavefront sensor in the adapter and an additional wavefront sensor in a test camera (Noethe and Guisard 2000).

A fast tip-tilt control of the secondary mirror with a bandwidth of a few Hz facilitates, in addition to chopping, also a second, fast stage for the control of the image motion.

Corrected modes and correction rate. The number of modes that have to be corrected can be defined by the method described in [▶ Sect. 4.4.2](#). The third mode of rotational symmetry 2, $e_{2,3}$, is the stiffest elastic mode that random forces with an rms of 2% of the nominal force of 1,500 N at zenith can generate with a coefficient larger than 15 nm. Therefore, apart from the modes $e_{0,1}$ and $e_{1,1}$, which are replaced by Zernike defocus and third-order coma, respectively, the set of active modes includes all the modes in [▶ Table 5-2](#) with eigenfrequencies up to 246 Hz, plus the next softest mode, $e_{0,3}$.

With six rings, even the highest active mode $e_{0,3}$ can be corrected with an accuracy of 99% ([▶ Fig. 5-16](#)). The required correction rate can be estimated from the variation of the coefficient of the most sensitive modes $e_{2,1}$ and defocus. According to [▶ Fig. 5-10b](#), the rms of the wavefront error generated by a variation of these coefficients over one minute is of the order of 100 nm. Therefore, the specification of 50 nm can be fulfilled with cycle times of 30 s.

The specification of 100 nm rms for the average effect of random force errors required an accuracy of the force setting for individual supports better than 0.6 N rms. Errors in the differential movements of the secondary mirror in any direction of the order of 5 μm and in the rotation about its center of curvature of approximately 0.3 arcsec generate wavefront errors of approximately 35 nm for defocus and 3 nm for third-order coma. Together with third-order astigmatism, defocus is therefore the most difficult mode to control.

Switch between foci. Each of the four telescopes is equipped with four focal stations, two Nasmyth stations, one Cassegrain, and one Coudé station. Since these have different back focal distances, the optical quality of the telescope could only be optimized for one focal station. The Ritchey-Chrétien condition is therefore only fulfilled at the Nasmyth focus.

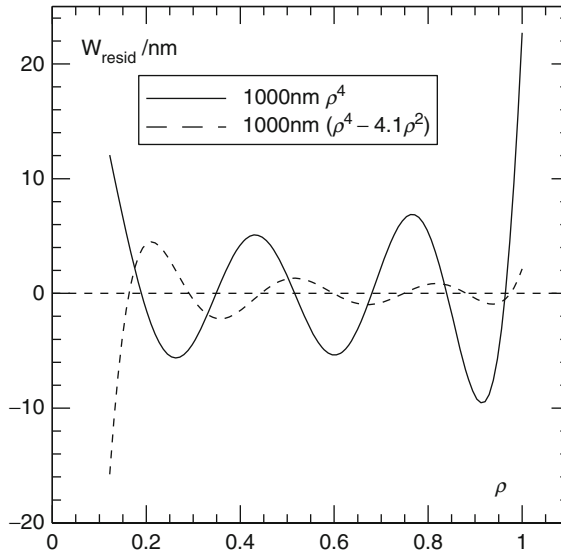
After refocusing, but without any other active optics corrections, the Cassegrain focus would suffer from field-independent spherical aberration ρ^4 with a coefficient of 20.8 μm as well as from third-order coma with a linear field dependence. Since the asphericity of the secondary mirror could not be modified, a correction with the only available degree of freedom, namely, a deformation of the primary mirror, could only eliminate the spherical aberration.

The best possible correction of 20.8 μm ρ^4 with six rings would leave a residual wavefront error w_{resid} with six nodes along the radius (solid line in [▶ Fig. 5-19](#)) and a residual rms of only 9 nm. However, the correction can be done even more efficiently both in terms of the residual error and the required forces if a defocus term $\alpha\rho^2$ is added to the pure spherical aberration ρ^4 . This produces a deformation that is much closer to the lowest elastic mode of rotational symmetry 0 and can therefore be generated with smaller forces and also more precisely than a pure ρ^4 deformation. The optimum value for α is -4.1 . Compared with the correction of pure ρ^4 , the residual error with an rms of the wavefront error of 2 nm is 4.5 times smaller ([▶ Fig. 5-19](#)), and the maximum forces of -180 N and $+470$ N are 45% smaller.

This also shows that the range of the active forces is primarily defined by the change of the configuration, rather than the correction of errors introduced by polishing or incorrect passive forces.

Uncorrectable aberrations. Special care has to be taken to minimize wavefront aberrations that cannot be corrected by the active optics system. One of these aberrations is the mirror seeing, which is generated when there is a temperature difference between the primary mirror and the ambient air. Therefore, the temperature in the enclosure is actively controlled by a ventilation system and the temperature of the primary mirror by a cold plate behind its back surface. During the day, the two devices set the temperatures of the structure, the mirrors, and the air in the enclosure to the temperature of the ambient air expected at the beginning of the night.

During the night, the cold plate is also used to equilibrate the mirror temperature with the falling temperature of the ambient air. Most of the time, the temperature differences between



■ Fig. 5-19

Residual wavefront aberration w_{resid} as a function of the normalized radial coordinate after optimum correction of spherical aberration ρ^4 (solid line) and $\rho^4 - 4.1\rho^2$ (dashed line)

the mirror and the ambient air can then be kept within a small range of $\pm 1^\circ\text{C}$. For such small temperature differences, the generated mirror seeing is insignificant (Cullum and Spyromilio 2000; Guisard et al. 2000).

Also the disturbances generated by the wind contain frequencies that are too high to be controllable by the quasistatic control of the optical elements of the VLT. Inside the enclosure, average wind speeds across the primary mirror of 1 m/s are associated with pressure fluctuations with an rms of the order of 1 N. These can cause deflections of the primary mirror in the form of the lowest elastic mode $e_{2,1}$ with rms values of the global wavefront error of approximately 150 nm. The errors scale with the square of the pressure variations. Therefore, under strong wind conditions, the primary mirror must be protected by an enclosure that is also equipped with a wind screen.

Another source of errors with frequencies ranging to well above those of wind pressure is vibration. Since the generated tip-tilt errors are difficult if not impossible to correct even by adaptive optics systems, vibrations must be avoided as much as possible by a proper design of several of the telescope components.

5.2.2 Open-Loop Performance of the VLT

Correction of initial aberrations. After the initial installation of the mirrors and the application of the nominal forces under the primary mirror, the two dominant aberrations with coefficients of about $20 \mu\text{m}$ were Zernike third-order coma and third-order astigmatism. They generated point spread functions with a FWHM of at least 2 arcsec.

They also generated a Shack-Hartmann pattern that was too heavily distorted to be analyzed by the standard algorithm in the wavefront analysis. However, an initial coarse correction, which was based on a visual inspection of strongly defocused images, removed a large fraction of the aberrations. The coefficients of the two dominant aberrations could be estimated from certain characteristics of the defocused images. In combination with a defocus, third-order astigmatism causes an elongation of the image, which changes its orientation by 90° when the defocus goes through zero. For third-order coma, the characteristic feature is a shift of the central obscuration with respect to the outer edge.

After a coarse correction of the two dominant aberrations, the Shack-Hartmann pattern was sufficiently regular to be usable for automatic analyses.

Hysteresis effects. Even with a precise knowledge of the calibrations for the rigid-body positions of the mirrors and the active forces under the primary mirror, the corrections could still be affected by hysteresis. To measure such effects, the zenith angle was modified in the range from 5° to 65° in steps of 15° several times up and down, and, at each position, the appropriate calibrated position and force corrections were applied. The wavefront measurements showed noticeable hysteresis of the order of 600 nm, 90 nm, and 15 nm for the elastic modes $e_{2,1}$, $e_{3,1}$, and $e_{4,1}$, respectively. The ratios of these values are in reasonably good agreement with the ratios of the inverse stiffnesses of the modes.

The mode $e_{2,1}$ with a coefficient of 600 nm generates a point spread function with a FWHM of approximately 0.25 arcsec. If this had been the only error and all other corrections based on calibrations had been perfect, the VLT could have been operated in open loop, at least under average seeing conditions. In practice, however, failures of mechanical components like an individual support would require a time-consuming setup of new calibration tables.

5.2.3 Closed-Loop Performance of the VLT

Minimum number of active modes. Over an altitude range of 70° , the rms values of the variations of the coefficients of the four stiffest active modes are of the order of only 10 nm (► [Fig. 5-10b](#)). Even the variations of the modes 8–12 are all smaller than 30 nm, and the variations of the modes 4, 6, and 7 are smaller than 100 nm. Therefore, for a seeing-limited performance, it would be sufficient to control, after an initial correction of all active modes, only the modes 1, 2, 3 and 5, that is, defocus, third-order coma and the lowest elastic modes of the rotational symmetries 2 and 3 during operation.

Residual aberrations. As in ► [Fig. 5-15b](#), the circles in ► [Fig. 5-20](#) show the rms of the variations of the low-order coefficients. However, this time, the coefficients were measured during operation with continuous closed-loop corrections of all active modes. If significant errors were introduced by the telescope mechanics, they would predominantly appear in the lowest elastic mode $e_{2,1}$. In comparison with the other coefficients, the one of $e_{2,1}$ should then be larger than expected from Kolmogoroff statistics. ► [Figure 5-20](#), which looks very similar to ► [Fig. 5-15b](#), shows that this is not the case. Mechanically induced errors, including those accumulating between two corrections, are therefore negligible compared with the ones introduced by the atmosphere.

As discussed before, the measured coefficients contain two components, a telescope component due to the errors in the telescope optics and an atmospheric component due to the effects of the atmosphere. On the one hand, without corrections, the telescope components change slowly. The correlations between successive values of the quickly varying atmospheric

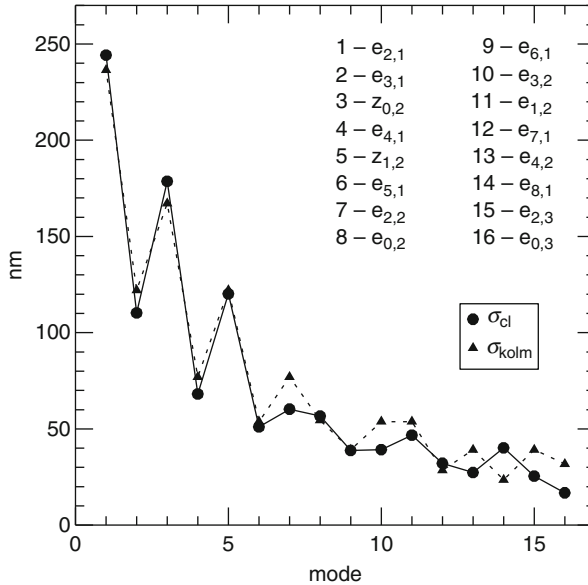


Fig. 5-20

Correlations between expected rms values of the variations of the coefficients due to Kolmogoroff turbulence (σ_{kolm}) and the rms σ_{cl} of the coefficients in closed-loop operation

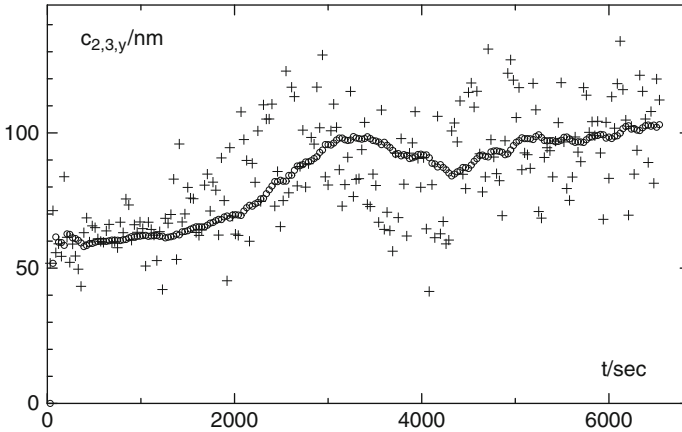
components that are due to the atmosphere are approximately 25%, which means that they are only weakly correlated.

On the other hand, in closed-loop active optics, the slow variations of the telescope optics will be largely corrected. However, the measured quickly varying atmospheric component will now be introduced into the telescope optics by the resultant correction. Consequently, it will appear as the major contributor to the telescope component in the subsequent measurement.

Therefore, the two components are of the same origin and of the same order of magnitude. Since they are largely uncorrelated, the rms values of the fast variations of the coefficients of the active modes measured in closed loop and shown in Fig. 5-20 should then be, and actually are, approximately $\sqrt{2}$ times larger than the corresponding rms values in Fig. 5-15b.

Filtering of the data. The noise introduced by the free atmosphere on the wavefront measurements cannot be avoided. However, the results can be filtered such that the active optics system corrects effectively only the slowly evolving errors introduced by the mechanics of the telescope. This can be accomplished very efficiently by a Kálmán filter, which is a special type of Bayesian filter. For each coefficient, the filter requires a priori knowledge about three parameters: first, the expected initial correction, say, after a preset; second, the expected systematic variation during a given time interval; and third, the expected noise.

All this information can be obtained from the results of the drift measurements summarized in the Figs. 5-10b and 5-15b. From Fig. 5-10b, the first parameter can be estimated from the differences of the coefficients for zenith distances typical for presets and the second parameter from differences evolving over time intervals of 30 s. The third parameter is directly given in Fig. 5-15b. The solid line in Fig. 5-21 shows how the coefficient of one of the components of the elastic mode $e_{2,3}$ would effectively follow its systematic evolution if the noise



■ Fig. 5-21

Measured coefficients of the y-component of the elastic mode $e_{2,3}$ with an added constant offset of 50 nm (dots) and the evolution of the coefficient, which would be corrected after filtering (solid line)

due to the atmosphere was filtered out. To demonstrate that the corrections converge quickly also for large initial errors, a constant offset of 50 nm has been added to the measured data.

An important consequence of the use of a filter is a significant reduction of the applied correction forces during each correction cycle. For the highest active modes like $e_{2,3}$ or $e_{0,3}$, the systematic variation over one minute is of the order of 1 nm, whereas the noise is of the order of 20 nm. The maximum force required for a correction of such a single, relatively rigid mode with a coefficient of 20 nm is of the order of 10 N. Consequently, the largest forces required for a correction of all active modes based on unfiltered coefficients are usually of the order of 30 N.

These large forces could be reduced to approximately 2.5 N if only the lowest three active elastic modes $e_{2,1}$, $e_{3,1}$, and $e_{4,1}$ were corrected. However, using the filtered data for the corrections, the maximum forces for the correction of the whole set of active modes are even smaller, namely, about 1.5 N.

Best performance of the VLT. Unlike the NTT, the VLT can achieve its specified image quality only with a two-stage control of the tracking. Using only the main axes drives, the residual tracking error was of the order of 0.1 arcsec rms. However, additional corrections by the secondary mirror with a bandwidth of 5 Hz reduced the tracking error to approximately 0.02 arcsec rms.

With continuous active optics corrections during the integrations, including the application of the Kálmán filter, the best recorded image quality was 0.25 arcsec FWHM for long exposure times of 5 min. For short exposure times of only a few seconds, the best measured value was 0.18 arcsec FWHM, which is close to the best seeing values ever measured on Paranal with devices dedicated to seeing measurements. These results prove that a large active telescope with mirror diameters of 8 m can achieve seeing-limited performance also under excellent seeing conditions. To reach even diffraction-limited image qualities requires, in addition, adaptive optics techniques, that is, also the correction of the wavefront aberrations generated by the atmosphere.

5.2.4 Other Large Active Telescopes

Telescopes with thin meniscus mirrors. Other 8 m telescopes that are similar to the VLT are the two Gemini telescopes and the Subaru telescope (Iye 1991). All of them are equipped with VLT-like thin meniscus mirrors with diameters of approximately 8 m. In the Subaru telescope, each axial support also acts as a lateral support. To avoid additional torques due to the lateral supports, the contact points of the supports are located in the neutral surface of the mirror, which required bores in the solid meniscus mirror.

Telescopes with structured mirrors. With the same weight, structured mirrors are considerably stiffer than solid ones. Nevertheless, all large structured mirrors such as the ones in the MMT and the Magellan Telescope with diameters of 6.5 m and the ones in the Large Binocular Telescope and the Large Synoptic Survey Telescope with diameters of 8.4 m are actively controlled (Martin et al. 1998, 2004; Schechter et al. 2002). Like the push-pull axial forces, also the lateral forces are applied at the back surface of the honeycomb mirror, which generates a global moment. The deformations introduced by this moment, as well as other low-order deformations, are corrected by 160 actuators under the 8.4-m mirrors, of which 108 actuators also apply controllable lateral forces. In addition, since the substrate is borosilicate glass with a nonnegligible coefficient of thermal expansion, a ventilation system minimizes the temperature gradients within the blank and also, to prevent mirror seeing, keeps its temperature close to the one of the ambient air.

The mirror is held in position by fixed points, which ideally take no loads. Any nonzero loads, which are detected by load cells, will be reduced by adjusting the active forces at a rate of 1 Hz. More details about these structured mirrors are given in ► Chap. 4 of this book.

The Large Synoptic Survey Telescope (LSST) will be an anastigmatic three-mirror survey telescope with a field of view of 3.5° (Krabbendam et al. 2010; Olivier et al. 2008). The primary and tertiary mirrors share the same blank, with the inner edge of the primary mirror coinciding with the outer edge of the tertiary mirror. Since the latter has a smaller radius of curvature, excess material that is left after the spin casting above the prescribed surface of the tertiary mirror has to be ground off before polishing. The convex secondary mirror is a solid meniscus with a diameter of 3.5 m and a thickness of 100 mm and is made of ultra low expansion glass. A special feature is that not only its axial but also its lateral support is active.

Reflective active Schmidt telescope. The use of active optics has also pushed up the size limit for wide-field telescopes based on the Schmidt design. The largest Schmidt telescope with a refractive aspheric corrector plate has a diameter of the entrance pupil, that is, of the Schmidt plate, of 1.34 m. However, larger Schmidt telescopes can be built with all-reflective optics (Lemaitre 1976).

The first telescope of this kind is the LAMOST telescope, a survey telescope with a field of view of 5° . The entrance pupil is a flat steerable siderostat of elliptical shape with a diameter of approximately 4 m along the short axis, which sends the light to a fixed spherical mirror with a diameter of approximately 6 m (Su et al. 1998). The two LAMOST mirrors consist of 24 and 36 hexagonal segments with diameters of 1.1 m corner to corner.

Similar to the aspheric plate in conventional Schmidt telescopes, the siderostat mirror has to correct the spherical aberration produced by the spherical mirror. Therefore, the surface of the siderostat mirror has to deviate from the prescribed flat shape, whereby the exact form depends on the pointing of the mirror. The change of the surface figure of the siderostat mirror is accomplished by active modifications of both the alignment and the shapes of the segments.

Segmented mirror telescopes. The thickness of the primary mirror can be substantially reduced by splitting the mirror into segments. The first large telescope of this kind was the Keck telescope with a diameter of the primary mirror of 10 m (Wizinowich et al. 1994). A detailed description of telescopes with segmented mirrors is given in [Chap. 3](#) of this book.

5.3 Future Extremely Large Telescopes

Segmentation of the pupil. Currently, the only feasible method to increase the collecting area of optical telescopes above the areas given by existing 8–10 m telescopes is the segmentation of, at least, the primary mirror. Broadly speaking, there are two approaches to segment a large mirror. In one approach, the segments fill, apart from small gaps, the aperture completely. In this case, the shapes of the segments are usually regular hexagons with diameters of the order of 1–2 m. An alternative would be to use a pattern with petals, where the borders of the segments would be along concentric rings and along the radii. The other approach to segmentation, which is pursued in the design of the proposed Giant Magellan Telescope, is to fill the aperture only partially with circular monolithic mirrors with diameters of several meters (Shectman and Johns 2010).

Two proposed telescopes use hexagonal segments of the primary mirror to fill the aperture, the Thirty Meter Telescope (TMT) and the European Extremely Large Telescope (E-ELT) with a diameter of the primary mirror close to 40 m.

Thirty Meter Telescope. The primary mirror of the TMT with a diameter of 30 m consists of approximately 500 hexagonal segments with diameters of 1.4 m corner to corner (Stepp 2012). The optical design is a Ritchey-Chrétien design with two Nasmyth foci. The control of the primary mirror will be very similar to the control of the primary mirror of the Keck Telescope. For the secondary mirror, a thin meniscus with a diameter of 3 m and a thickness of 100 mm, there are two options. It can be an active mirror that controls the lowest modes of rotational symmetries 2, 3, and 1 with approximately 60 supports on four rings. Alternatively, it can be a passive mirror since all aberrations generated by the secondary mirror that are constant across the field can, with high accuracy, be corrected by the primary mirror.

European Extremely Large Telescope. The E-ELT (Gilmuzzi and Spyromilio 2007) with a diameter of the primary close to 40 m marks, in terms of control, a further step in the evolution of telescopes, namely, from active to adaptive telescopes. In addition to measuring and correcting devices already used in active telescopes, an adaptive mirror forms an integral part of the E-ELT. The adaptive correction features are essential to fulfill the major performance requirements on the telescope.

Specifications for the E-ELT. First, over a field of 5 arcmin, the telescope by itself should correct the wavefront aberrations generated by the ground layer of the air. This alone will not be sufficient to reach a diffraction-limited performance. However, the improvement of the image quality will be equivalent to a reduction of the seeing by a factor of approximately 2, which can also be expressed by the metric $\text{PSSN} = 2$ as defined in [Sect. 2.2](#).

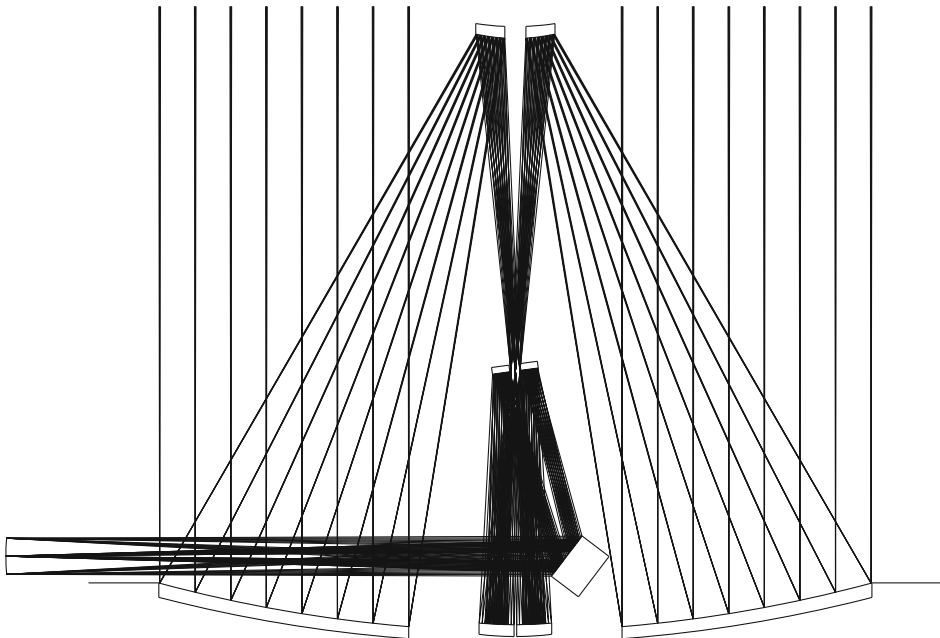
Second, over a small field of view of only a few arcseconds and under seeing conditions of 0.7 arcsec FWHM, the telescope, again by itself, should be capable of delivering diffraction-limited images at wavelengths of $2.2 \mu\text{m}$ with a Strehl ratio of 0.7. For this performance, the rms of the total wavefront error, which includes the residual aberrations due to the atmosphere after adaptive optics corrections, must not be larger than 200 nm.

Another specification is a diffraction-limited performance at wavelengths of $2.2\ \mu\text{m}$ with a Strehl ratio of 0.7 over a field of at least 20 arcsec, which can only be achieved with an additional adaptive mirror in the instrument. Used in tandem, the two adaptive mirrors can simultaneously correct wavefront errors generated in the ground layer and at high altitudes.

Optical Design of the E-ELT. Since the adaptive mirror has to be conjugated to a layer a few hundred meters above the ground, it is optically nearly conjugated to the primary mirror and can therefore not be used for the corrections of field aberrations. The third requirement, asking for diffraction-limited optics over a fairly large field, which can only be fulfilled by an anastigmatic optical design, therefore requires a third aspheric mirror sufficiently far from the pupil.

For technological reasons, the correction of the image motion is split between two mirrors. One of them corrects only medium-amplitude image motion at medium frequencies. The other one, the adaptive mirror, corrects both the residual small amplitude image motions and the wavefront shape errors at high frequencies.

❶ *Figure 5-22* shows the anastigmatic optical design of the E-ELT with three powered mirrors and two flat mirrors (Delabre 2008). The primary mirror with an elliptic figure consists of approximately 800 segments, the hyperbolic secondary mirror is a convex solid meniscus with approximately 100 active pneumatic supports, and the third mirror is a weak general asphere with approximately 70 active pneumatic supports. M4 is the flat adaptive mirror with approximately 4,000 actuators and a thickness of 2 mm, and M5 the flat mirror for medium-amplitude, medium-frequency tip-tilt corrections.



■ Fig. 5-22

Optical design of the adaptive European Extremely Large Telescope

Adaptive and active optics operation of the E-ELT. For its control, the E-ELT requires four wavefront sensors. Three of them sample the wavefront with approximately 4,000 subapertures at a rate of about 700 Hz. Their reference stars form, depending on the availability of sufficiently bright stars, as closely as possible a triangle in the field. The fourth sensor measures the segment figures and the phase steps for the calibration of the edge sensors. This can be done at intervals of a few weeks if the stability of the edge sensors and the segment support system is sufficient to rely on infrequently updated calibrations. Otherwise, if closed-loop corrections are required, the sensor can deliver the correction signals at rates of approximately 0.01 Hz.

Whereas active and adaptive optics are usually separated, this is no longer the case in the E-ELT. All wavefront errors are first and nearly instantaneously corrected by the adaptive mirror M4 across the full field. Before M4 runs out of its correction range, the accumulated errors will be offloaded at lower frequencies to the other active elements.

However, with three wavefront sensors and assuming that field distortions cannot be measured, a total of about 20 error sources cannot be disentangled with sufficient accuracy. For example, an appropriate reshaping and simultaneous realigning of the segments can generate nearly perfect third-order astigmatism across M1. Independently, M2 and M3 can also generate third-order astigmatism. Without the possibility to measure distortion in the field, the three contributions cannot be separated and only their sum can be measured.

Multistage control of tip-tilt errors. Because of the large errors and the high temporal frequencies to be corrected, the control of the E-ELT is not quasistatic anymore, but highly dynamic. For example, the lowest eigenfrequency of the E-ELT is of the order of 2–3 Hz, which limits the bandwidth for the correction of tracking errors with the main drives to effectively 1 Hz. The residual tracking errors after the first stage would then be approximately 0.3 Hz, which is barely sufficient even for a seeing-limited performance.

In a second stage, corrections with M5 with a bandwidth of 5 Hz can reduce the tracking errors to approximately 0.04 Hz. In a third stage, these residual errors will then be further corrected by the adaptive fourth mirror with a correction bandwidth of 12 Hz to residual errors of approximately 0.002 arcsec.

Observations of exoplanets require an extreme diffraction-limited performance, which can, in a fourth stage, be reached with another adaptive mirror in the instrument. Apart from residual wavefront shape errors, this mirror corrects the tracking errors at a bandwidth of approximately 120 Hz and reduces them to 0.0005 arcsec. Such cascaded control loops will also be used for the correction of other aberrations.

Complex control will be essential to extend the active optics operations, which were sufficient to reach seeing-limited performances in 8-m class telescopes, to adaptive optics operations, which will enable diffraction-limited observations in even larger telescopes.

Acknowledgments

The authors would like to thank Steffan Lewis, Andrew Rakich, Jason Spyromilio, and Isabelle Surdej for helpful discussion and suggestions.

References

- Bely, P. Y. ed., 2002, *The Design and Construction of Large Optical Telescopes* (Berlin: Springer)
- Born, M., & Wolf, E. 1997, *Principles of Optics* (6th ed.; Oxford: Pergamon Press)
- Couder, A. 1931, *Bull Astron*, 2me Série, Tome VII, Fasc. VI, 243
- Creedon, J. F., & Lindgren, A. G. 1970, *Automatica*, 6, 643
- Cullum, M., & Spyromilio, J. 2000, *Proc SPIE* 4004, 194
- Delabre, B. 2008, *Astron Astrophys*, 487, 389
- Dierickx, P. 1992, *J Mod Opt*, 39, 569
- Dierickx, P. 1994, *Proc SPIE* 2199, 950
- Diffrient, R. 1994, *Sky Telesc*, 91
- Fried, D. L. 1965, *J Opt Soc Am*, 55, 1427
- Gerchberg, R. W., & Saxton, W. O. 1972, *Optik*, 35, 237
- Gilmozzi, R., & Spyromilio, J. 2007, *The Messenger*, 127, 11
- Gitton, P., & Noethe, L. 1998, *The Messenger*, 92, 15
- Gonsalves, R. A. 1976, *J Opt Soc Am*, 66, 961
- Gonsalves, R. A. 1982, *Opt Eng*, 21, 829
- Gonsalves, R. A., & Childlaw, R. 1979, *Proc SPIE* 207, 32
- Guisard, S., Noethe, L., & Spyromilio, J. 2000, *Proc SPIE*, 4003, 154
- Hubin, N., & Noethe, L. 1993, *Science*, 262, 1390
- Hull, T., et al. 2003, *Proc SPIE*, 277
- Iye, M. 1991, *JNLT Technical Report No. 2*
- Krabbendam, V.L., Sweeney, D. and the LSST Collaboration 2010, *Proc SPIE* 7733, 77330D
- Lassell, W. 1842, *Mem R Astron Soc* XII, 265
- Lemaitre, G. 1976, *J Opt Soc Am*, 66, 1334
- Lemaitre, G. 2009, *Astronomical Optics and Elasticity Theory* (Berlin: Springer)
- Lukosz, W. 1963, *Opt Acta*, 10, 1
- Maksutov, D. D. 1954, *Technologie der astronomischen Optik* (Berlin: VEB Verlag Technik), 74
- Martin, H. M., et al. 1998, *Proc SPIE*, 3352, 412
- Martin, H. M., Cuerden, B., Dettmann, L. R., & Hill, J. M. 2004, *Proc SPIE* 5489, 826
- McLeod, B. A. 1996, *PASP* 108, 217
- Noethe, L. 1991, *J Mod Opt*, 38, 1043
- Noethe, L. 2001, *Active Optics in Large Telescopes with Thin Meniscus Primary Mirrors*, Habilitationsschrift (Technische Universität Berlin)
- Noethe, L. 2002, *Program Optim*, 43, 1
- Noethe, L. 2005, *J Mod Opt*, 52, 603
- Noethe, L. 2009, *Exp Astron*, 26, 1
- Noethe, L., & Guisard, S. 2000, *Astron Astrophys. Suppl. Ser.* 144, 157
- Noethe, L., & Guisard, S. 2000, *Proc SPIE* 2003, 382
- Noll, R. J. 1976, *J Opt Soc Am*, 66, 207
- Olivier, S.S., Seppala, L., Gilmore K. and the LSST camera team 2008, *Proc SPIE* 7018, 70182G
- Piatrou, P., & Chanan, G. 2010, *Appl Opt*, 49, 6395
- Platt, B., & Shack, R. V. 1971, *Opt Sci Center News* (University of Arizona, Tucson), 5(1), 15
- Racine, R., Salmon, D., Cowley, D., & Sovka, J. 1991, *Proc Astron Soc Pac*, 103, 1020
- Raggazoni, R. 1996, *J Mod Opt*, 43, 289
- Ray, F. B. 1991, *Proc SPIE* 1532, 188
- Roddi, F. 1988, *Appl Opt*, 27(7), 1223
- Roddi, F. 1999, *Adaptive Optics in Astronomy* (Cambridge/New York: Cambridge University Press)
- Sarazin, M., Melnick, J., Navarrete, J., Lombardi, G. 2008, *The Messenger*, 132, 11
- Schechter, P. L., & Sobel, R. E. 2010, *Publ Astron Soc Pac*, 123(905), 812
- Schechter, P. L., et al. 2002, *Proc SPIE*, 4837, 619
- Schneermann, M., Cui, X., Enard, D., Noethe, L., & Postema, H., 1990, *Proc SPIE* 1236, 920
- Schwesinger, G. 1988, *J Mod Opt*, 35, 1117
- Schwesinger, G. 1991, *J Mod Opt*, 38, 1507
- Schwesinger, G. 1994, *Appl Opt*, 33, 1198
- Seo, B. J., et al. 2009, *Appl Opt* 48, 5997
- Shack, R. V., & Platt, B. C. 1971, *JOSA*, 61, 656
- Shack, R. V., & Thompson, K. 1980, *Proc SPIE*, 251, 146
- Shectman, S., & Johns, M. 2010, *Proc SPIE* 7733, 77331Y, doi: 10.1117/12.857852
- Stanghellini, S., Legrand, P., Baty, A., & Hovsepian, T. 1997, *Proc SPIE* 2871, 314
- Stepp, L., 2012, *Proc SPIE* 8444, 84441G, doi: 10.1117/12.9280006
- Su, D., Cui, X., Wang, Y., & Yao, Z. 1998, *SPIE*, 3352, 76
- Tango, W. J. 1977, *Appl Phys* 13, 327
- Thompson, K. P. 2005, *J Opt Soc Am A*, 22, 1389
- Wilson, R. N. 1982, *Opt Acta*, 29, 985
- Wilson, R. N. 1999, *Reflecting Telescope Optics II* (Berlin: Springer)
- Wilson, R. N. 2004, *Reflecting Telescope Optics I* (Berlin: Springer)
- Wilson, R. N., Franza, F., & Noethe, L. 1987, *J Mod Opt*, 34, 485
- Wilson, R. N., Franza, F., Noethe, L., & Andreoni, G. 1991, *J Mod Opt*, 38, 219
- Wizinowich, P., Mast, T., Nelson, J., & DiVittorio, M. 1994, *Proc SPIE*, 2199, 94

6 **Optical and Infrared Interferometers**

*Theo A. ten Brummelaar*¹ · *Harold A. McAlister*²

¹Mount Wilson Observatory, The CHARA Array of Georgia State University, Mount Wilson, CA, USA

²Department of Physics and Astronomy, Georgia State University, Atlanta, GA, USA

1	<i>Introduction</i>	243
2	<i>Basic Theory</i>	243
3	<i>The Effect of the Atmosphere</i>	247
4	<i>Visibility Calibration and Closure Phase</i>	249
5	<i>Basic Design</i>	251
6	<i>Beam Combination</i>	266
	<i>References</i>	280

Abstract: Stellar interferometers achieve limiting angular resolution inaccessible to even next-generation single-aperture telescopes. Arrays of small or modest apertures have achieved baselines exceeding 300 m producing submilliarcsecond resolutions at visible and near-infrared wavelengths. The technical cost and challenge in building interferometric arrays is substantial due to the very high tolerance imposed by optical physics on the precision of beam combination and optical path length matching for two or more telescopes. This chapter presents the basic theory and overall design considerations for an interferometer with an emphasis on the practical aspects of constructing a working instrument that overcomes obstacles imposed by the atmosphere, submicron path length matching requirements, limitations on number of telescopes and their layout, light losses through multiple reflections and transmissions necessary to superimpose telescope beams in the beam-combining laboratory, and other realities of the art of interferometry. The basic design considerations for an interferometer are laid out starting with site selection and telescope placement and then followed through to beam combination and measurement of interferometric visibility and closure phase after the encountering of numerous subsystems by incoming wavefronts. These subsystems include active wavefront sensing for tip/tilt correction or even full-up adaptive optics, telescope design for directing collimated beams over large distances, diffraction losses, polarization matching, optical path length insertion and active compensation, correction for atmospheric refraction and differential dispersion in glass and air, separation of light into visible and near-infrared channels, alignment over long optical paths, high-precision definition of the three-dimensional layout of an interferometric array, and, finally, a variety of beam-combining schemes from simple two-way combiners to multitelescope imaging combiners in the pupil and image planes. Much has been learned from a modest but robust collection of successful interferometers over the last 25 years or so, and interferometry is poised to become a mainstream technique for astrophysical research.

Keywords: Adaptive optics, Adaptive optics: sampling time, Advantages/disadvantages, Alignment system: strategies, Aperture plane combination: Michelson interferometer, Array layout: UV coverage, Astrometric model: three-dimensional array layout, Atmospheric phase error, Atmospheric seeing, Baseline, Baseline solution, Beam expansion, Beam reduction, Bispectrum, Bootstrapping, Calibration, Calibrator object, Cat's-eye, Closure phase, Construction stability and vibration: vibration isolation, Control systems, Control tolerance, Correlation, Coupling losses, Delay line, Dichroics, Differential diffraction and dispersion: angular refraction, Diffraction: beam propagation, Dither mirrors, Diversity of beam combiners, Fiber-based beam combiners, Field of view: baseline dependency, Fixed delay, Fizeau interferometry, Fourier transform, Fringe amplitude, Fringe detection strategies, Fringe envelope, Fringe tracker, Fringe tracking: group delay tracking, Geometric delay, High angular resolution, Hilbert transform, Image plane combination: spectrographic combination, Image reconstruction, Inner and outer scale lengths, Instrumentation, Integrated optics beam combiners, Interferometer design, Laser metrology, Light pipes: vacuum propagation, Longitudinal dispersion, Minimizing optical surfaces, Opposing glass wedges, Path length control: time variable delay, Optical fibers and spatial filtering: coherence loss with aperture, Optical path length difference, Optical path length equalizer, Optical quality and throughput: light losses, Optical windows, Phase locking, Phase variance, Piston error, Polarimetry: polarization state separation, Polychromatic fringe, POP mirrors, Quantification of seeing, Polarization: reflection sequencing, Quantification of turbulence, r_0 , Remote control, Reynold's number, Risley prisms, Separating fringe tracking from science: detection strategies, Shutter sequences, Siderostat, Single-mode fibers, Site selection: criteria, Spatial filtering, Spatial fringe encoding, Speckle

interferometry, Strehl ratio, τ_o , Taylor hypothesis, Telescope arrays, Telescope placement, Telescopes: collimated beams, Temporal fringe encoding, Thermal stability, Tip/tilt control, Triple correlation, Turbulent flow, UV plane, Visibility amplitude, Visibility from fringe power spectrum, Visibility function, Visibility phase, Wollaston prism, Young's double slit, Zernike/Van Cittert theorem

1 Introduction

Optical interferometry is a technique that brings the light of many telescopes together in a single location in order to create very high angular resolution images. These images have the resolution of a telescope with a diameter the size of the largest separation between the smaller telescopes, albeit with much less sensitivity. An optical interferometer is a large and complex instrument, comprised of many active and passive subsystems involving many engineering challenges in their design and construction. These include, but are not confined to, physical stability, atmospheric seeing, path length equalization, beam combining, fringe tracking, dynamic control, and dispersion/diffraction considerations. In this chapter some of these critical aspects will be discussed along with the solutions that the community has used to date.

A seminal overview of what is required to build an interferometer is given by Tango and Twiss (1980). This is a “must read” for any student of the subject. Other excellent starting points for reading are the book by Labeyrie et al. (2006) and the course notes of the 1999 Michelson Interferometry Summer School (Lawson 1999). More recent articles on the practice of Interferometry can be found in *New Astronomy Reviews* volumes 51 and 53. Good reviews of relatively recent astrophysics done using this technique can be found in Monnier (2003) and Quirrenbach (2001), while all the important historical papers on the subject are collected together in Lawson (1997). For the most recent news and publications, the Optical Long Baseline Interferometry Newsletter (olbin.jpl.nasa.gov) is an excellent resource.

2 Basic Theory

In its simplest form, an interferometer brings together the light from two telescopes and mixes them. The electric fields of the two beams from the two telescopes can be written as

$$\begin{aligned} A_1 &= I_1 \cos(kx_1 + kct), \\ A_2 &= I_2 \cos(kx_2 + kct), \end{aligned}$$

where I_1 and I_2 are the amplitudes, $k = \frac{2\pi}{\lambda}$ is the wave number, λ is the wavelength, x_1 and x_2 are the optical paths from the star through the two telescopes all the way to the point at which the two beams are combined, c is the speed of light, and t is the time. Once these two signals have been mixed, the camera detects the time-averaged intensity which will be given by

$$\begin{aligned} I &= \overline{(A_1 + A_2)^2} \\ &= \overline{I_1^2 \cos^2(kx_1 + kct) + 2I_1 I_2 \cos(kx_1 + kct) \cos(kx_2 + kct) + I_2^2 \cos^2(kx_2 + kct)} \\ &= \frac{(I_1^2 + I_2^2)}{2} \\ &\quad + 2I_1 I_2 [\overline{\cos(kx_1) \cos(kct) + \sin(kx_1) \sin(kct)}][\overline{\cos(kx_2) \cos(kct) + \sin(kx_2) \sin(kct)}] \end{aligned}$$

where the $\overline{\cos^2(kct)} = 0.5$ substitution has been used. This can be further simplified by noting that $\cos(kct) \sin(kct) = 0$, resulting in


$$\begin{aligned} I &= \frac{I_1^2 + I_2^2}{2} + I_1 I_2 [\cos(kx_1) \cos(kx_2) + \sin(kx_1) \sin(kx_2)] \\ &= \frac{I_1^2 + I_2^2}{2} + I_1 I_2 \cos(k[x_1 - x_2]). \end{aligned}$$

The detected signal is then an oscillating intensity pattern with a magnitude and a phase, called the *correlation* of the two waves. In most practical applications, this function is normalized by the mean intensity $(I_1^2 + I_2^2)/2$, resulting in

$$I(v) = 1.0 + T(I_1, I_2) V(v) \cos(2\pi v \text{OPD} + \phi(v)),$$

where $\text{OPD} = x_1 - x_2$ is the *optical path length difference*, and the wave number k has been replaced with $v = 1/\lambda$. The transfer function

$$T(I_1, I_2) = \frac{2I_1 I_2}{I_1^2 + I_2^2}$$

has also been introduced. This has a value of 1.0 if the two beams are of equal intensity. This visibility function can also be expressed as a complex number, because it has an amplitude and a phase, and is a function of the intensity pattern of the object being studied on the sky. The fringe phase is, in fact, the same as the phase difference between the two incoming wavefronts. This visibility function can also be expressed as a complex number, because it has an amplitude and a phase, and is a function of the intensity pattern of the object being studied on the sky. All of this is, in its essence, identical to the well-known Young's double slit experiment for monochromatic light as shown in  Fig. 6-1.

All of this has assumed monochromatic light, but in practice of course astronomical objects emit light at many wavelengths, and an optical filter, or spectrograph, of some kind is normally used to restrict the bandwidth to a known finite size, say from $v_0 - \Delta v/2$ to $v_0 + \Delta v/2$. This requires an integration over the entire pass band, which we here assume is flat and finite, resulting in

$$I(v_0, \Delta v) = \frac{1}{\Delta v} \int_{v_0 - \Delta v/2}^{v_0 + \Delta v/2} [1.0 + T(I_1, I_2) V(v) \cos(2\pi v \text{OPD} + \phi(v))] dv.$$

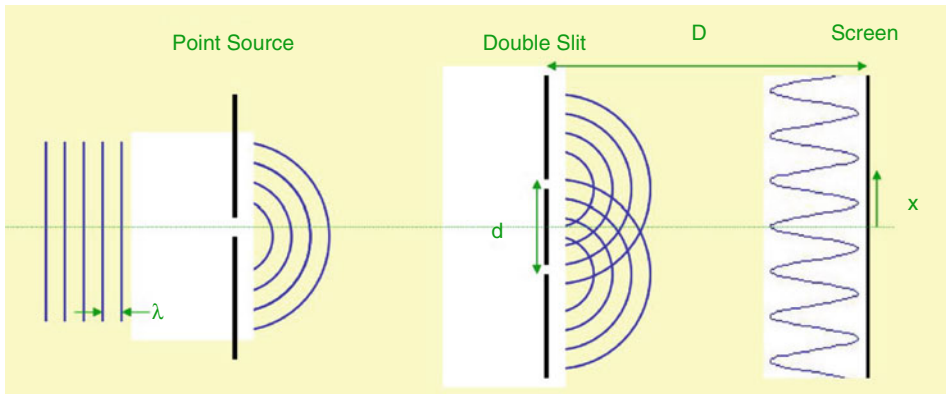
The first part of this integral is trivial, and the second part can be simplified by assuming that the visibility function does not change very much inside the small bandwidth. This leads to

$$I(v_0, \Delta v) = 1.0 + \frac{1}{\Delta v} T(I_1, I_2) V(v_0) \int_{v_0 - \Delta v/2}^{v_0 + \Delta v/2} \cos(2\pi v \text{OPD} + \phi(v)) dv. \quad (6.1)$$

Next, the substitution $\eta = v - v_0$ is made, and it is then relatively easy to show that

$$I(v_0, \Delta v) = 1.0 + T(I_1, I_2) V(v_0) \text{sinc}(\pi \text{OPD} \Delta v) \cos(2\pi v_0 \text{OPD} + \phi(v_0)). \quad (6.2)$$

So what was an infinitely large fringe pattern for a single wavelength becomes a finite *fringe packet* when a nonzero bandwidth is considered. The modulation of the fringe pattern, which in this case is a *sinc* function, is called the *fringe envelope* and has a maximum when $\text{OPD} = 0$ and its first zeros when $\pi \text{OPD} \Delta v = \pm \pi$. The full width of the fringe envelope, known as the *coherence length*, is therefore given by $2/\Delta v = 2\lambda^2/\Delta \lambda$. The effect of a finite band pass is demonstrated



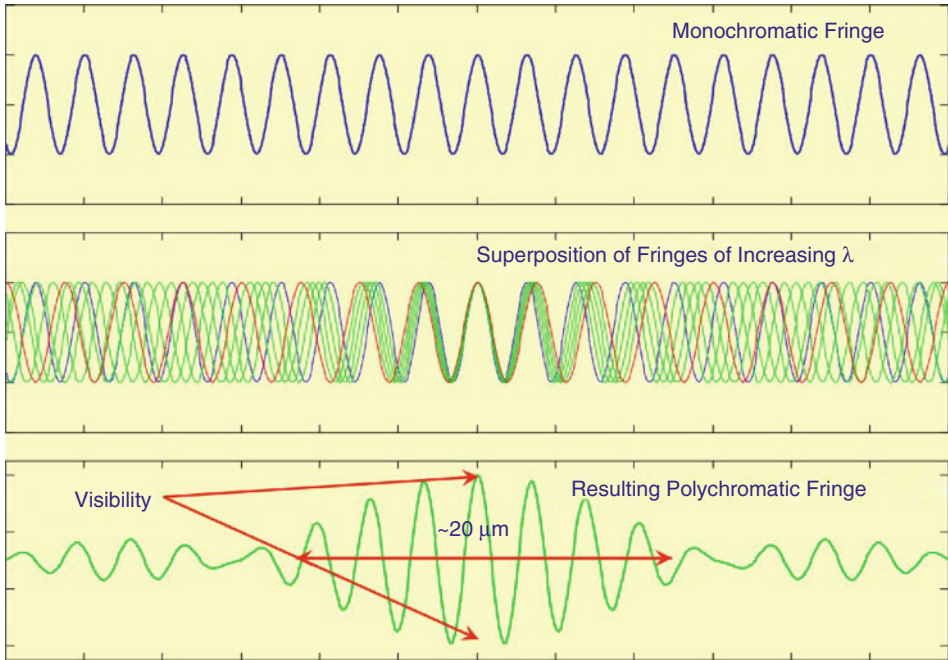
■ Fig. 6-1

Young's double slit experiment. Monochromatic light enters from the *left* through a single small aperture, which is then sent through two small apertures and onto a screen on the *right*. The result is a continuously oscillating intensity pattern of *fringes* on the *right*. Replacing the first small aperture with a star and the two next apertures with two telescopes pointing at that star turns this into a simple stellar interferometer. This can be done in practice by placing a piece of cardboard with two small holes over the aperture of a small telescope and using a very-high-power eyepiece to observe a star. The result will be an image of the star with fringes across it

graphically in ● Fig. 6-2. For example, a standard filter in the near infrared centered on $2.3\ \mu\text{m}$ and a width of $0.3\ \mu\text{m}$ will have a coherence length that is only $34\ \mu\text{m}$ wide. This means that the optical path length, from the star to the point at which the two beams are combined, must be controlled at the sub- μm level. This can impose very challenging constraints on the design of an optical stellar interferometer.

In this calculation, the band pass was assumed to be flat, that is, the band pass is a top-hat function, and it is no coincidence that the shape of the envelope, a *sinc* function, is the Fourier transform of a top-hat function. In actual practice, filters are not perfectly top-hat shaped and have rounded edges. Furthermore, the filter shape is convolved with the spectrum of the star as well as with the wavelength-dependent effects of the atmosphere and the optical system itself. In general, the fringe envelope shape is always the Fourier transform of the resulting convolved filter shape, which is a function of the filter used, the optical system, and the spectrum of the star. This relationship can be used to perform a kind of spectroscopy, known as *Fourier Transform Spectroscopy*, or FTS (Davis et al. 2001).

● Equation 6.2 is a simplified version of the fringe equation, and it is central to the art of stellar interferometry, both from scientific and engineering points of view. The scientific implications derive from the Zernike/Van Cittert theorem (Van Cittert 1934; Zernike 1938) which states that the correlation of the radiation collected by two separate apertures or telescopes is one Fourier component of the intensity pattern on the sky. The spatial frequency of this Fourier component is a function of the distance between the two telescopes, called the *baseline*, divided by the mean wavelength of the band-pass filter λ_0 . This Fourier plane is called the *UV plane*, and the parameters *U* and *V* are used to designate the two orthogonal spatial frequency axes. If the amplitude and phase of the visibility function can be measured for many baselines and many orientations on the sky, an inverse Fourier transform of these measurements in the *UV*



■ Fig. 6-2

While monochromatic light produces fringes of an infinite extent, multiple wavelengths, when combined together, reduce the extent of the fringes to a very small range of optical path length differences, often as small as a few tens of microns. This can impose very tight tolerances on the path length control systems in a stellar interferometer

plane can be used to produce an image of the object of interest, which has a spatial resolution that is a ratio of the wavelength of light being used and the largest baseline, or telescope separation. For a 100-m baseline with $\nu_0 = 0.5 \mu\text{m}$, this is of the order of 1 milliarcsecond. While it is not yet possible to build a telescope with a 100-m diameter, it is possible, indeed it has been done many times, to place two smaller telescopes 100 m apart and bring their light together in a central location. This is the reason why interferometers are built.

One way of looking at the Zernike/van Cittert theory is to compare an interferometer to a single large lens. A lens is itself a Fourier transform machine (See, e.g., Goodman 2005), that is, the intensity pattern in the image plane is the Fourier transform of the wavefront reaching the aperture. Thus a point source, like a star, will produce a flat wavefront at the aperture of the telescope if atmospheric distortions are not present. If this aperture is circular, the incoming wavefront will be a flat round disk, and the Fourier transform of a flat disk is the Airy disk pattern so often seen in a simple optical system. An interferometer can be seen as an instrument for sampling the incoming wavefront presented to a very large input aperture, and instead of using a lens to perform the transformation from aperture plane to image plane, it is done using image reconstruction software. This does assume, however, that enough amplitudes and phases are measured to make image reconstruction possible. A complete derivation of the Zernike/van Cittert theorem can be found in Born and Wolf (2002).

3 The Effect of the Atmosphere

Of all the sources of noise, the atmosphere is the most significant contributor. The atmosphere is in an unpredictable and complex state of turbulent flow, and an understanding of the theory of optical propagation through a turbulent atmosphere is important because it is the characteristics of the atmosphere that determine many of the design constraints of a ground-based interferometer, indeed of any ground-based astronomical system. There is only enough space for a very brief introduction here, but more detailed reviews of atmospheric turbulence theory and how it affects optical propagation can be found in Roddier (1981) and Coulman (1985).

The usual approach, first formulated by Reynolds, is to describe these flows using ensemble averages rather than in terms of individual components. He defined a nondimensional quantity, now known as the Reynolds number, that characterizes a turbulent flow. It is defined as

$$R = \frac{UL}{\nu_{\text{mol}}},$$

where U and L are the typical velocity and length for the flow and ν_{mol} is the kinematic molecular viscosity. A low Reynolds number indicates that the flow is *laminar*, that is, regular and smooth in space and time, while a large Reynolds number signifies a highly turbulent flow. Between these extremes, the fluid will undergo a series of unstable states. The atmosphere is difficult to study because it has a Reynolds number of the order of 10^6 . In the classical theory, proposed by Lev Landau, the number of the unstable states between laminar and fully turbulent flow would be very large, even infinite. More recent work in the area of chaos theory shows, however, that the final state of full-blown turbulence can arise after only a few such transitions.

The standard model for atmospheric turbulence and its effect on light propagation, first published by Taylor (1921) and Richardson (1922) and later expanded by Taylor (1935) and Kolmogorov (1941a, b), states that energy enters the flow of the atmosphere at low spatial frequencies as a direct result of the nonlinearity of the Navier-Stokes equation governing fluid motion. This forms eddies of large sizes which have a length L_0 known as the outer scale length. This outer scale length varies according to local conditions, ranging from the distance to the nearest physical boundary when close to the ground up to the thickness of the largest turbulent eddies. Measurements of L_0 range from 2 m (Nightingale and Buscher 1991), through to 40–60 m (Davis et al. 1995) all the way out to the rather controversial figure of 2 km (Colavita et al. 1987).

These large eddies are unstable and break up into smaller eddies, corresponding to a different scale length and a higher spatial frequency. These “second-generation” eddies are also unstable and will break up into still smaller eddies. Since the scale length associated with these eddies decreases, the Reynolds number associated with the flow must also be decreasing. When the Reynolds number is low enough, the turbulent breakup of the eddies stops, and the kinetic energy of the flow is lost as heat via viscous dissipation. This imposes a highest possible spatial frequency on the flow beyond which little or no energy is available to support turbulence. This inner scale length is denoted as l_0 . The inner scale length varies from a few millimeters near the ground up to about 1 cm high in the atmosphere. Richardson (1922) neatly described this turbulent cascade in the couplet:

*Big whirls have little whirls that feed on their velocity,
And little whirls have lesser whirls and so on to viscosity.*

However, simply stating that the atmosphere is turbulent does not imply it will affect optical propagation. It is possible to have a fluid in which strong mechanical turbulence will not affect optical propagation, for example, an incompressible fluid-like water. It is changes in the refractive index of the air, and not turbulence in itself, that cause changes in optical propagation. Since the refractive index of air is a function of its density and turbulence changes the density, there is a direct link between air turbulence and optical propagation.

The inner and outer scale lengths of the atmosphere determine many of the properties of atmospheric seeing. For example, the inner scale length is almost always smaller than the diameter of the telescopes, and there is therefore no escaping the fact that these turbulent cells will be blown past the aperture and cause distortions in the wavefront. Furthermore, the outer scale length is often of the same order as the distance between telescopes. So long as the outer scale length is larger than the telescope separation, the affects of the atmosphere on the observations will increase as the baseline is increased. Once the baseline is larger than the outer scale length, this will no longer be true and the detrimental effects of the atmosphere will get no worse with increasing telescope separation.

In order to establish bandwidths for the various servo systems in an interferometer, the temporal properties of the wavefronts entering our telescopes must be understood. This is done by first studying the spatial properties and then moving to the temporal domain by using the *Taylor hypothesis of frozen turbulence*. The Taylor hypothesis means that a frozen piece of turbulence is blown past our apertures by the prevailing wind. It is called frozen because it is assumed that it does not change shape significantly during the time it takes to move past the telescope aperture. The temporal characteristics of the wavefront entering the telescope will therefore be a function of the spatial distribution of turbulent cells, the speed of the prevailing wind, and the size of the aperture or baseline.

The atmospheric affects on optical propagation are characterized by defining two parameters: the atmospheric coherence length r_0 , which should not be confused with the fringe coherence length, and the atmospheric coherence time τ_0 . The atmospheric coherence length r_0 , where $l_0 < r_0 < L_0$, is defined as the length over which the wavefront can be well approximated by a flat wavefront, and τ_0 is defined as the time over which that assumption is valid, where our definition of “flat” is having an rms phase error of π radians. One way of looking at this is to think of r_0 as the size of the “typical” turbulent cells being blown past the aperture of a telescope, or between telescopes, and that these cells are being blown past by the wind at a velocity of r_0/τ_0 , a manifestation of the Taylor hypothesis. Another way of looking at this is that r_0 is the largest diameter a telescope can have and still be diffraction limited. A “typical” value for r_0 is 10 cm, which corresponds to 1-arcsecond seeing. If the wind were moving at 10 ms^{-1} , it would result in a characteristic time scale of $\tau_0 = 0.1/10 = 0.01 \text{ s}$ or 10 ms. Time scales in an interferometer range from the value of τ_0 up to the time it takes for the largest L_0 -sized cells to be blown between the telescopes, which can be many tens of seconds.

A single large aperture telescope contains many r_0 -sized areas. These can be thought of as small patches of parallel wavefronts spread across the aperture, each of which producing a diffraction-limited image in a different part of the image plane. As the prevailing wind blows these past the aperture, these multiple diffraction-limited images dance around in the image plane on a time scale of τ_0 . This is the origin of the well-known speckle patterns seen in high magnifications of the image plane of a large-aperture telescope. As first pointed out by Labeyrie (1974), the morphological information in these speckles can be used to obtain diffraction-limited astrometric data using the technique of *Speckle Interferometry*. Both r_0 and τ_0 scale

as $\lambda^{6/5}$, resulting in atmospheric seeing improving with observational wavelength. It is for this reason that a majority of interferometry is done in the infrared.

The value of τ_0 is in the range of a few milliseconds in the visible during normal seeing conditions and can be as high as several tens of milliseconds in excellent seeing in the infrared. Without a full adaptive optics system on each telescope, this basically sets the time constant for the servo systems and detector systems within the interferometer, that is, it will be necessary to sample the incoming wavefronts on time scales of order τ_0 . The value of r_0 can range from several centimeters in the visible to many tens of centimeters in the near infrared. As for the coherence time, the size of the coherence length will place engineering tolerances on the design of the interferometer itself. It turns out that the optimum size of telescope aperture without adaptive optics is of order $3r_0$ (Buscher 1988) and so in the near infrared, where r_0 is of order 10–40 cm in good seeing telescopes with apertures of 1–2 m are used. In order to use larger telescopes, full adaptive optics systems for each telescope would be required.

It is, of course, important that the seeing internal to the instrument does not further reduce the visibility. To this end, it is common to keep as much as possible of the light paths within the instrument in a vacuum, and in most cases, evacuated tubes are used to carry the light from the telescopes to the beam combination area. Once the light reaches the delay line and beam combination building, there are several other things that can be done to reduce the internal seeing, most importantly ensuring that the thermal properties of the beam combination facility are well known and controlled.

4 Visibility Calibration and Closure Phase

It is impossible to know the true value of the change in visibility amplitude introduced by the combination of atmospheric and instrumental effects. What is certain, however, is that both the atmosphere and all but the more obscure instrumental effects will lower the fringe amplitude by a random amount. If it is assumed that the amount of visibility reduction is, on average, constant over relatively short time periods and within nearby offset locations on the sky, this effect can be calibrated out by observing objects whose visibility is well known. Ideally this would be a nearby and unresolved single round star whose visibility will always be 1.0. If the visibility amplitude of the calibrator is found to be $V_{\text{cal}}(\nu)$ and the measured fringe amplitude of the object of interest is $V_{\text{obj}}(\nu)$ then

$$V(\nu) = \frac{V_{\text{obj}}(\nu)}{V_{\text{cal}}(\nu)}.$$

In practice, of course, it is not always possible to find such a star close in angular separation to the object of interest, and resolved objects of a well-known size must be used. Fortunately, for objects only partially resolved, the visibility amplitude is not a strong function of size and the errors introduced by the uncertainty in calibrator size do not affect the final calibrated visibilities very much. Furthermore, $V_{\text{cal}}(\nu)$ is a time-varying function, especially in poor seeing, and the calibrator must be measured both before and after the object of interest and then interpolated between these two measurements. Visibility amplitude calibration is one of the more difficult aspects of the analysis of interferometric data, and bad choices in calibrators, or poor observation method, have often led to poor, or even erroneous, results.

Apart from reducing the visibility amplitude, the atmosphere, along with some instrumental effects, changes the optical path lengths and introduces random visibility phase errors. While

it is always possible to measure the fringe phase when the atmosphere is present, this is not a measure of the Fourier phase. Without good phase measurements, an inverse Fourier transform cannot be performed. These atmospherically induced phase errors totally overwhelm the real phase of the visibility function, and on a single-baseline instrument, the true phase of the coherence function cannot be determined at all. Fortunately, for instruments with more than two telescopes, there is a way around this problem. While these phase errors are a function of the atmosphere above each telescope and are associated with the individual telescopes, the visibility function phase is a differential measurement between two telescopes, and this relationship can be used in the following way. Consider three telescopes ($i = 1, 2, 3$) where the true phase of the incoming light above the atmosphere is ϑ_i so that the visibility phase signal between any two telescopes is $\phi_{ij} = \vartheta_i - \vartheta_j$. Note that this visibility phase is the same as $\phi(\nu)$ in the fringe equation (6.2) but the dependence on the baseline has been added and the dependence on wave number ν has been removed for clarity. At each telescope, the atmosphere, as well as the instrument itself, introduces a random phase error σ_i . For each baseline, or pair of telescopes, the measured phase difference will be

$$\psi_{ij} = (\phi_i + \sigma_i) - (\phi_j + \sigma_j) = \phi_{ij} + \sigma_i - \sigma_j.$$

In most cases the random atmospherically induced phase errors $\sigma_i - \sigma_j$ totally overwhelm the visibility phase ϕ_{ij} . This means that the phase of the fringes of a single baseline is not readily observable and will almost always average out to zero. If instead three telescopes are used, the sum of fringe phases around a closed triangle of baselines will be

$$\begin{aligned} \Psi_{123} &= \Psi_{12} + \Psi_{23} + \Psi_{31} \\ &= \phi_{12} + \sigma_1 - \sigma_2 + \phi_{23} + \sigma_2 - \sigma_3 + \phi_{31} + \sigma_3 - \sigma_1 \\ &= \phi_{12} + \phi_{23} + \phi_{31}. \end{aligned}$$

Note that the random phase errors cancel out. This quantity, the sum of phases on a closed triangle of baselines, is called the *closure phase* and is a good measurable in the sense that many measurements of the closure phase will average out to the sum of the true visibility phases despite the phase errors introduced at each telescope by the atmosphere. Unlike the visibility amplitude, no calibration is required, although in a real instrument there will always be some small offsets in the closure phase that need to be removed. These offsets are constant and do not vary in time, or on the sky, and are a function of the beam combiner hardware, not the atmosphere. They will therefore appear as a constant, nonzero closure phase signal in the calibrator stars that can be subtracted from the closure phase signal of the object.

So while the phase information for each individual baseline is lost, enough phase information is available in the closure phases to perform the inversion from Fourier space and create the required image (Monnier 2011). In its simplest sense, a closure phase is a measurement of the nonsymmetry of the morphology of the object. A circularly symmetric object will have a closure phase of zero, while a binary star will have a very large closure phase.

It is extremely important here to note that the triangle must be truly closed, that is, for example, ψ_{31} and not ψ_{13} must be used, otherwise the random errors will not cancel and this will no longer be a good measurable. Furthermore, in practice, the complex visibility

$$V_{ij}(\nu) = V_{ij}(\nu)e^{i\phi_{ij}(\nu)}$$

is normally used for each baseline to calculate the visibility triple product, also known as the bispectrum or triple correlation (Lohmann et al. 1983):

$$\begin{aligned} V_{123}(\nu) &= V_{12}(\nu)V_{23}(\nu)V_{31}(\nu) \\ &= V_{12}(\nu)V_{23}(\nu)V_{31}(\nu)e^{i[\phi_{12}(\nu)+\phi_{23}(\nu)+\phi_{31}(\nu)]}. \end{aligned}$$

The mean of this triple product over many fringe samples can then be used to calculate a mean closure phase. Since the fringe visibility varies constantly from one sample to the next, this results in a mean closure phase weighted by fringe amplitude, a better value than the more direct mean of the closure phase.

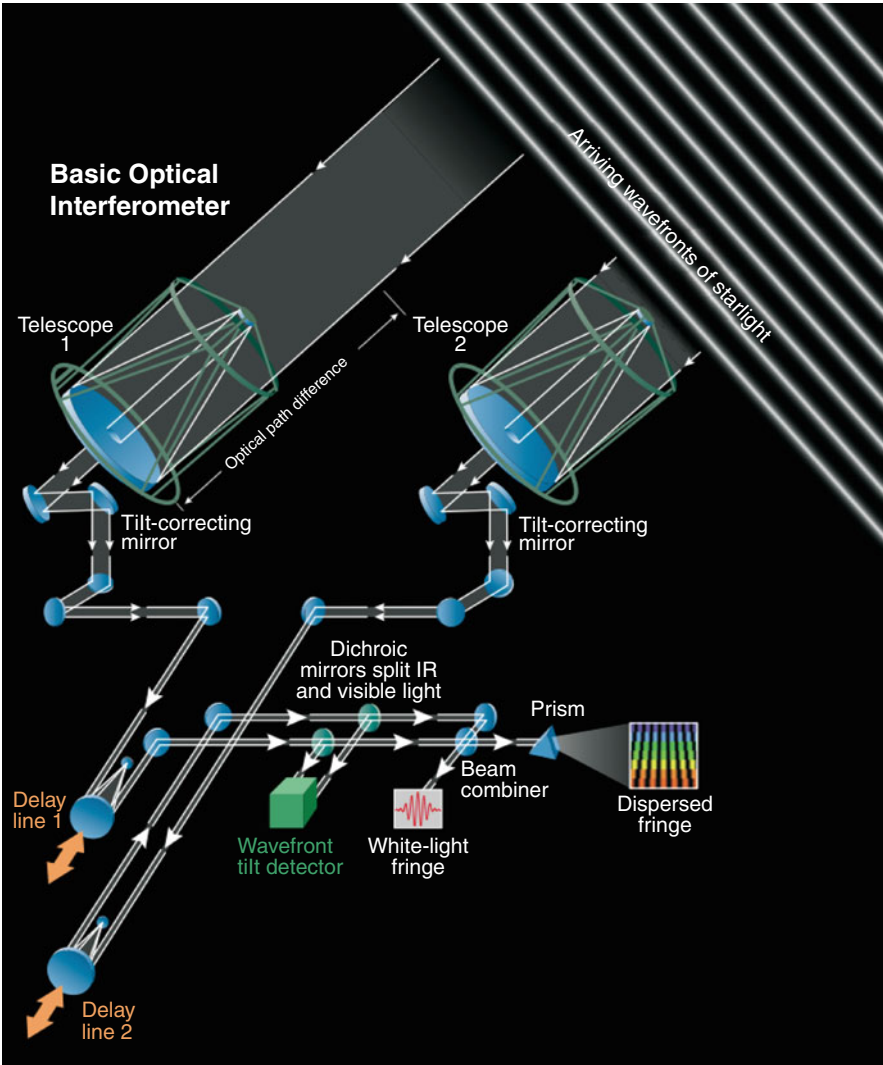
Interferometry, then, is in practice a matter of measuring as many visibility amplitudes and closure phases as possible and either fitting a model to these data or performing the inversion from Fourier space to image space and creating a picture. In the remainder of this chapter, the various hardware and control systems required to collect these data will be described.

5 Basic Design

The basic layout of a two-dimensional single-baseline interferometer is shown in [Fig. 6-3](#). A geometric delay of $B \sin \theta$, where B is the baseline length and θ is the zenith angle of the object, is created whenever the telescopes are pointing at a star that is not exactly at zenith. In a real three-dimensional system, one uses the dot product of the baseline vector and the unit vector pointing at the star to determine the geometric delay. Furthermore this delay is continually changing as the earth rotates and, as discussed above, continually modulated by the atmosphere, and even by optical systems within the interferometer itself. Since the fringe envelope is only a few tens of microns across, it is necessary to ensure that the optical paths, from star to fringe position, are the same. This is done by measuring the path difference between the two beams and then adding or subtracting optical path in each beam using an optical system called a *delay line* or *optical path length equalizer* (OPLE) to make the path difference zero. This requirement of path equalization is one of the most difficult aspects of interferometry and imposes some of the strongest constraints on the design of the optical system. We note that because we are dealing with optical frequencies, the technology does not currently exist to time encode the incoming signals and combine beams *a posteriori* in software as has long been utilized in astronomical radio interferometry.

There are numerous other active optical systems required. For example, atmospheric distortions of the wavefront reaching each telescope need to be controlled using either a tip/tilt servo or a full adaptive optics system. Additionally, differential air path lengths within the instrument introduce differential longitudinal dispersion, which also may need to be corrected. Here each of the important subsystems will be described in the order that they are normally seen as the wavefront moves through the instrument.

There are numerous design decisions to be made when building an interferometer, and, of course, no two groups make exactly the same choices. Many of these choices will be strongly dependent on the available funding, the main scientific thrust of the research group or groups involved, and the availability of existing infrastructure. [Table 6-1](#) shows a list of existing ground-based facilities and a very short summary of the basic features of each. For more information, refer to the citations listed.



■ Fig. 6-3

Basic layout of a single-baseline interferometer. Two telescopes point at the same object, then send the light through a tilt correction system, an optical path length equalizer, and finally to a central location where fringes are formed. The most important problem to be solved is ensuring the light paths, from the star all the way to the fringe position, are equal to within a small fraction of the coherence length of the light (figure courtesy of Sky and Telescope magazine and Dr. Peter Lawson)

Site Selection – Selecting a site for an interferometer is very much like site selection for any astronomical observatory but with the added constraint of requiring a large relatively flat area in order to accommodate the large baselines required. The fact that there needs to be an unobscured line of site between each telescope and the beam-combining facility must also be taken into account. The largest possible baseline will be determined by the morphology of the

Table 6-1

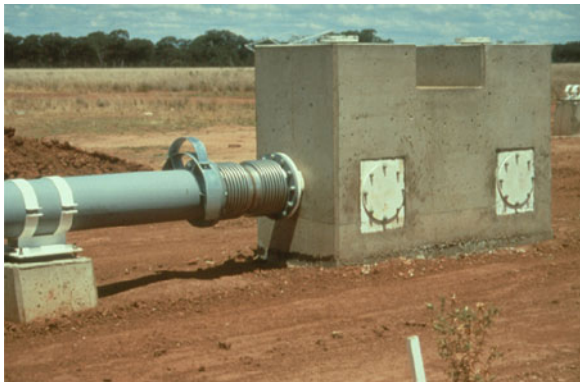
This table contains a very brief overview of existing and under construction facilities at the time of writing of this chapter listed in order of their commissioning. Items in brackets are upgrades expected in the near future and the expected dates of first results. Note that the LBTI is unlike the rest of those listed as it uses a combination of filled aperture and interferometry that is beyond the scope of this chapter. Also, the ISI uses a heterodyne combination method similar to that used by radio interferometers that is also beyond the scope of this chapter. Here, visible light means a wavelength of 0.5–1 μm , near infrared is 1–3 μm , and mid-infrared is around 10 μm

Name	Citation	Apertures	#Combined	Baselines	Wavelengths	Location	Status
SUSI	Davis et al. (1998)	13 × 15 cm fixed	2	5–160 m (640 m)	Visible	Narrabri, Australia: low-altitude flat plain	Operational
ISI	Hale et al. (2000)	3 × 165 cm movable	3	5–80 m	Mid-infrared	Mount Wilson, CA, USA: mountaintop	Operational
NPOI	Armstrong et al. (1998)	6 × 12 cm movable	6	7–79 m (437 m)	Near infrared	Anderson Mesa, AZ, USA: mesa	Operational
Keck	Colavita et al. (2003)	2 × 10 m fixed	2	85 m	Near and mid-infrared	Mauna Kea, HI, USA: mountaintop	Operational
CHARA	ten Brummelaar et al. (2005)	6 × 1 m fixed	6	34–341	Visible and near infrared	Mount Wilson, CA, USA: mountaintop	Operational
VLT	Haguenauer et al. (2010)	4 × 8.2 m fixed 4 × 1.8 m Movable	4	11–129 m	Near infrared	Cerro Paranal, Chile: mountaintop	Operational
LBTI	Hinz et al. (2004)	2 × 8.4 m fixed	2	0–23 m	Near and mid-infrared	Mount Graham, AZ, USA: mountaintop	Under construction (2011)
MROI	Creech-Eakman et al. (2010)	10 × 1.4 movable	6	7–340 m	Visible and near infrared	South Baldy, NM, USA: mesa	Under construction (2012)

observatory site, and ideally a large flat area at the top of a mountain is desired. Unfortunately this is hard to find, as mountaintops tend to have complex rather than flat terrain. Existing interferometers (See ▶ [Table 6-1.](#)) have been built in a variety of locations, with some on traditional mountaintops and others on planes located much closer to sea level. In the case of the VLTI, the top of the mountain was removed in order to create a large flat area at high altitude. Other inland interferometers are placed on mesas, which by their very nature are large, flat, and at high altitudes. Most, however, are on more traditional mountaintops and must deal with the local topography. Ultimately, of course, the choice will be a compromise of some sort, and will also include factors such as existing infrastructure, weather patterns, and proximity to nearby towns and cities.

Construction Stability and Vibration – Since an interferometer has engineering tolerances of the same order as the wavelength at which it will operate, everything must either have an active optical system for correction or be stable at the submicron level. Not only are large inertial masses required but it is also necessary to isolate them from any local sources of vibration. Thus, the foundations of the buildings must not be physically coupled to the telescope support systems, delay lines, or optical table supports. Sometimes this is impossible to do, and vibrations from the building and other instruments and nearby equipment are transferred to the interferometer optics. In these cases, it is necessary to measure these vibrations with either accelerometers, or with the *fringe tracker* described below and remove them by sending the appropriate error signals to the delay lines. In the end, this requirement means interferometers contain a great deal of concrete and steel, as can be seen in ▶ [Fig. 6-4.](#)

The buildings must also have very good thermal stability; otherwise, internal paths and optical alignment will continually change with temperature. For the telescopes and domes, it is only necessary to use the standard methods, such as removal of local electronic heat sources, of ensuring minimal dome seeing. The problem of the necessarily large beam-combining facility is different. It is not wise to have moving air from an air conditioning system inside the optical laboratory, as this will create very turbulent air and a large amount of internal seeing. The two most common solutions are to either place the facility under ground or use a “building within



■ Fig. 6-4

One of the concrete piers that support the siderostats for the SUSI interferometer. The concrete goes down into the ground all the way to local bed rock to ensure very high stability and low levels of vibration (Figure courtesy of Professor John Davis)

a building” where the laboratory is inside the inner building, which contains a large but passive thermal mass, and the space between the two buildings serves as an HVAC plenum volume.

Array Layout – The positioning of the input apertures, be they siderostats or telescopes, is often a case of logistics and is always a matter of solving several, often conflicting, requirements. As many telescopes as possible are desired as well as a large range of baseline sizes and orientations. In order to enable a “snap-shot” mode, that is, have the ability to collect enough data in a short time to form an image, the minimum number of telescopes is normally said to be six. This is largely a case of how many baseline sizes and orientations can be collected at any one time, something referred to as *UV coverage*. As with the choice of site, the available funding often has a large influence over the number of telescopes, as does the topography of the site itself. One way of getting around this is to have telescopes that can be moved around, but this, of course, adds its own complexities.

Another consideration in array layout is whether or not to have a nonredundant baseline arrangement. All optical interferometers have a limited number of telescopes, and so in most cases they are arranged such that the distance and orientation of any two is unique. This gives the maximum number of different baselines for a given number of telescopes. However, one of the most difficult problems to overcome in an interferometer is locating and tracking the fringes, that is, controlling the delay lines in order to keep the fringes in a stable position within the beam combiner. Invariably, the larger a baseline is, the lower the visibility amplitude will be and, depending on the baseline size and current outer scale length of the atmosphere, the more the fringes will move around. This makes finding and tracking fringes on very large baselines extremely difficult. If there are enough telescopes, it is possible to get around this by using a technique called *bootstrapping* where the telescopes are placed relatively close together in a long, usually redundant, line of telescopes. The separation of each pair of telescopes is then small, and finding and tracking the fringes between each local telescope pair is relatively easy. If each adjacent pair of telescopes is phased, then all the telescopes in that line are in phase, and it is then certain that all the other baselines, including the longest, are also correctly phased.

Telescopes – The input apertures of an optical interferometer need to be able to direct the stellar light toward the central facility, and so the output of these systems, unlike imaging telescopes, will almost always be a collimated beam of light. In some cases, and for smaller apertures, no powered optics are required at all and a flat mirror in an alt/az mount can be used. This is known as a *siderostat*, similar to the heliostats used in many solar observatories. For larger apertures, even if siderostats are used, powered optics are needed in order to bring the beam diameter down to a manageable size. In the case of a telescope utilizing confocal paraboloids for the primary and secondary optics, maintenance of a perfectly collimated beam from the telescope into the beam-combining area at the center of the interferometer array requires the use of temperature-insensitive structures, such as Invar rods, to maintain a nearly constant distance between the primary and secondary.

The design of the telescopes, or siderostats, will also be influenced by the need for the structural stability discussed above, and this places constraints on these subsystems not normally found in standard telescopes. For example, longitudinal path length modulations and vibrations do not adversely affect single-aperture imaging telescopes but can be a major concern in an interferometer. The result is that telescopes designed for interferometers normally are much more massive than single-aperture systems, and the control electronics and other potential sources of heat and vibration are frequently placed in a separate building from the telescopes themselves.

Since, by definition, an interferometer will include multiple telescopes spread out over a large area, possibly kilometers across, each telescope must be designed as a stand-alone system capable of complete remote control, including mirror covers, fiducials, telescope pointing, dome control, and acquisition and finder cameras.


Adaptive Optics – The wavefronts reaching the telescopes have been distorted as they pass through the atmosphere, and some of this at least needs to be corrected, preferably as close to the telescope itself as possible. For smaller telescopes, a tip/tilt, or image stabilization, servo is enough (ten Brummelaar and Tango 1994) though a full adaptive optics system is of course much better (Tyson 2000). It is desirable to place this subsystem as close as possible to the telescope, for example, using the secondary mirror to act as a tip/tilt device is common. The wavefront and image position for each telescope must be measured, but it is not always clear where this sensor system should be. Since the optical system between the telescope and the beam-combining laboratory can also introduce wavefront tilt and higher order aberrations, it has until recently been standard practice to place the sensor system in the beam-combining laboratory itself. Unfortunately this also means that it is placed behind the many reflections in the optical system and will therefore lose sensitivity. To maximize sensitivity, the sensor system needs to be located as close to the telescope as possible. More recently it has become the practice to have a high-speed, that is, sampling as fast or faster than the atmospheric coherence time of τ_0 , tip/tilt or wavefront detector at the telescope to close the servo loop and a much slower detector system in the beam-combining laboratory that can detect any noncommon path errors introduced by the remaining optical system.

One useful bonus of having a tip/tilt system is that the motion of the mirror required to keep the stellar image stable is a direct measurement of the wavefront tilt imposed by the atmosphere. Thus, it is possible to extract an estimate of the atmospheric coherence length r_0 from these data. The variance of the angle of arrival σ_θ^2 as measured by the tip/tilt servo is directly related to the diameter of the telescope D and the coherence length r_0 (Noll 1976 and Tango and Twiss 1980) via

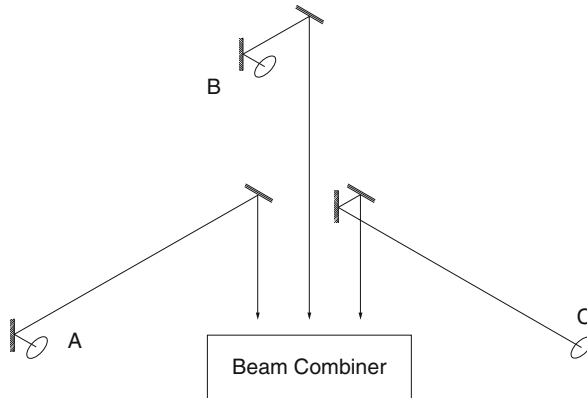
$$\sigma_\theta^2 = 0.184 \left(\frac{D}{r_0} \right)^{5/3} \left(\frac{\lambda_0}{D} \right)^2,$$

and this can be inverted in order to find an estimate for r_0 . Of course, if an AO system is used, much more sophisticated measurements of atmospheric turbulence are possible.

Polarization – Each reflection from a mirror surface introduces a phase shift between polarization states. This means that the reflections used in each arm of an interferometer must be the same: reflection symmetry should not be broken. Reflection symmetry is also important in order to have the same image rotation in each arm (Traub 1988). For a two-dimensional interferometric array, this means that there will be more reflections than are desirable, but there is little that can be done about this.

Two complications arise when designing a completely symmetric series of reflections to bring the light from the telescopes into the optics lab where the beams are combined. These are that the telescopes are spread out in a two-dimensional pattern, and that they may also be at different heights, that is the pattern may actually be three-dimensional. Mountaintops are rarely completely flat. One solution to the two-dimensional case is shown in  Fig. 6-5. A similar solution to the three-dimensional problem would require one more reflection in each telescope line.

Light Pipes – Once the telescopes have sampled the light, it must be sent to the beam-combining laboratory, and this is why the output beam of the telescopes is normally a collimated



■ Fig. 6-5

One method of preserving reflection symmetry in the optical system of an interferometer. Each telescope output beam undergoes the same reflections in the same order, thereby ensuring that the polarization states will be the same in the beam combiner. A similar solution exists for the three-dimensional problem but involves one more reflection per telescope line

rather than a diverging or converging beam. This will almost always be done close to the ground, and in order to avoid sending the beam through a great deal of ground-level air turbulence, it is commonly placed inside a vacuum tube system. This also has advantages in combating dispersion in the air as will be discussed below. A more recent trend is to use matched length single-mode optical fibers (Vergnole et al. 2005), although they are temperature sensitive, cause substantial losses, and a different set of fibers for each optical band is required.

Diffraction – Interferometers have, by their very nature, very long path lengths of many hundreds of meters. A collimated optical beam of this length will necessarily diffract, and this can have very detrimental effects. There are two main concerns with diffraction in an interferometer. First, the paths within the interferometer will not be the same for each telescope line and there will therefore be differential diffraction that can reduce the system visibility. This can in theory be calculated (Sheppard and Hrynevych 1992) and calibrated out, but this is difficult at best. Second, a lot of diffraction will result in the beam size expanding as the beam propagates, no matter how carefully it has been collimated at the telescope. One can actually take advantage of this to spatially filter the beams (Horton et al. 2001). The problem of diffraction can be avoided altogether by using optical fibers for beam transport, but they bring with them a number of other difficulties including dispersion, light loss, temperature sensitivity, and the need for a different fiber for each waveband of interest.

In the end, the amount of diffraction must be minimized and the amount of diffraction is proportional to the beam reduction factor squared and the propagation distance. While there is little that can be done about the propagation distance, which will be determined by the array geometry, the reduction factor must be chosen very carefully by selecting the maximum beam size practicable for the beam propagation and having a second beam reduction much closer to the final point of beam combination. This can only be taken so far, however, and as baselines of many kilometers come into existence, this solution will ultimately fail as the beam size required will become too large and unwieldy, making the beam transport system prohibitively expensive.

Another method of avoiding the problems of diffraction is to reimage the telescope pupil inside the beam-combining system, thereby ensuring that the wavefront reaching the beam combiner is the same as that sampled by the telescope primary mirror. Unfortunately, since the distance between the telescope primary mirror and the beam combiner is constantly changing (see the Path Length Control section below), this requires a variable focus device of some kind to continually ensure that the primary mirror is correctly imaged in the beam-combining laboratory by use of a variable curvature mirror or lens system of some kind. For example, the VLTI makes use of a variable curvature mirror controlled by compressed air (Ferrari et al. 2003).

Optical Quality and Throughput – Ultimately, it is the signal-to-noise ratio that will determine how faint the interferometer can go and how long it takes to reach the desired measurement precision. By their very nature, optical interferometers have a very large number of optical surfaces, both reflective and transmissive, and each of these surfaces degrades the wavefront as well as reduces the total number of photons reaching the beam-combining laboratory.

The most common form of measurement of visibility amplitude is $V^2(\nu)$, and if detector noise and background counts are, the signal-to-noise ratio is given by¹

$$\text{SNR}(V^2(\nu)) = N_{\text{PH}} V^2(\nu),$$

where N_{PH} is the number of detected photons. The number of surfaces depends on the details of the design, but it is not unusual to have of the order of 20 reflections from the telescope primary to the beam-combining laboratory. If each reflection is 90% efficient, the total throughput is only 12%. Optical coatings are rarely this good, however, and even less often stay that good over time. The upshot is it is extremely important to have as few reflections as possible, something very difficult to achieve given that the solutions to most of the other problems in designing an interferometer involve adding reflections. Using optical fibers to transport the light can reduce the total number of reflections, but as discussed above, they come with their own performance costs.

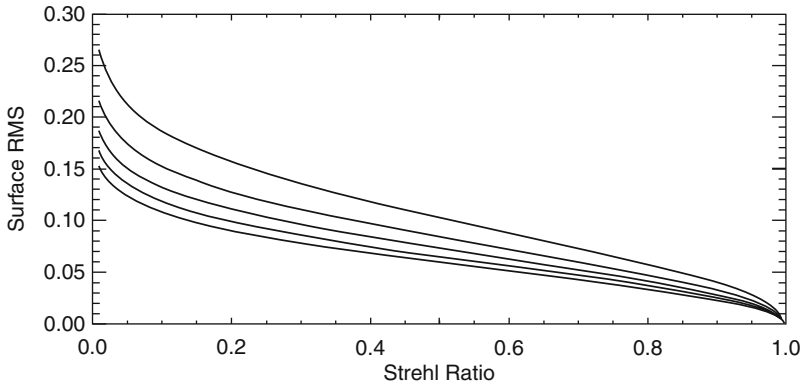
In order to characterize the affect of $V(\nu)$ on the signal-to-noise ratio, the way the optical quality of the components affects the final raw visibility must be investigated. The parameter most commonly used to characterize the quality of an astronomical system is the *Strehl ratio* $S(\nu)$ defined as the ratio of the intensity of light in the center of the image and the intensity of light in the center of a perfect image, most often an Airy disk. So a perfect optical system has a Strehl ratio of 1.0, while the atmosphere in good seeing conditions on a large telescope rarely has a Strehl ratio of above 0.01 in optical wavelengths. Like r_0 and τ_0 , $S(\nu)$ is a strong function of the seeing and observational wavelength and, in an interferometer, the visibility. A good approximation for the Strehl ratio is given by

$$S(\nu) \approx e^{-\sigma^2(\nu)},$$

where $\sigma^2(\nu)$ is the variance of the phase errors at the wave number ν and the phase variance is related to the peak to valley error of the optical component $E_{\text{PV}}(\nu)$ via

$$\sigma^2(\nu) \approx \left(\frac{2\pi E_{\text{PV}}(\nu)}{3.47} \right)^2.$$

¹A more detailed discussion of signal-to-noise ratio in the presence of detector noise can be found in ten Brummelaar 1996.



■ Fig. 6-6

The surface rms error required to achieve a given Strehl ratio for a range of numbers of reflections starting with 10 at the *top*, 30 at the *bottom*, and moving up in increments of 5 reflections

A more common way to specify optical surface quality is the root mean square surface error, or E_{RMS} , and a good approximation of this is $0.07E_{\text{PV}}(\nu)$. Combining these equations results in

$$E_{\text{RMS}}(\nu) \approx \frac{2.45}{2\pi} \sqrt{\frac{\ln S(\nu)}{-N_{\text{REF}}}},$$

where N_{REF} is the number of reflections in the optical system. This function is plotted in ● Fig. 6-6 for the full range of Strehl ratio and a number of reflections ranging from 10 to 30 in increments of 5.

It turns out that under good conditions, the Strehl ratio and the raw visibility have a one-to-one correspondence, that is, $V(\nu) \approx S(\nu)$ (ten Brummelaar et al. 1995). This means that the raw visibility can never be greater than the Strehl of the optical system. So, for example, the minimum raw visibility of the optical system is set to 0.9, the individual optics in the system all need to be $\lambda/20$ or better, and this is almost independent of the total number of reflections in the system. Optics up to about 2 in. in diameter of this quality are commercially available, but anything larger will need to be custom built. Furthermore, any powered optics, such as the telescopes and any other beam reducers in the system, need to achieve wavefronts as good or better than this.


Finally, the transmissive optics, such as windows on the vacuum system, dichroics, and lenses, need to be considered. As for the reflective optics, any flat or powered element like a lens will have to meet the tolerances discussed above. There are also other considerations for the flat transmissive optics. For example, while nonreflective coatings are extremely efficient, it is very difficult to make one equally efficient across the very wide optical bandwidths most interferometers use, and there will always be residual reflections on both surfaces of an optical window or dichroic. Also, if the two surfaces of these optics are parallel, there is a real risk of causing multiple internal reflections, so it is standard practice to purposefully introduce an angle between the two sides of the substrate, normally of the order of 20 seconds of arc. This in turn introduces a small amount of angular dispersion into the beam, which, over the very long propagation lengths and wide bandwidths in an interferometer, can be quite significant. This can be avoided by ensuring that the windows at either end of a vacuum system have matching and opposite

wedges wherever possible. This, however, does not correct for the wavelength-dependant beam offset caused by dispersive elements in the optical train, only the angular offset.

Path Length Control – Basic theory states that the typical envelope width is only a few tens of microns across, that there is a continually changing geometric delay between the telescopes as the earth rotates, and that the atmosphere is continually changing this path length difference. In fact, the internal paths of an interferometer must be controlled to a tolerance much smaller than the observation wavelength. In order to keep the fringe location stable, the location of the fringes must be known, which will be discussed in the following section, and a way of adding or subtracting optical path in each telescope line is also required. The hardware used to adjust internal path lengths is called the *optical path length equalizer* (OPLE) or *delay line*.

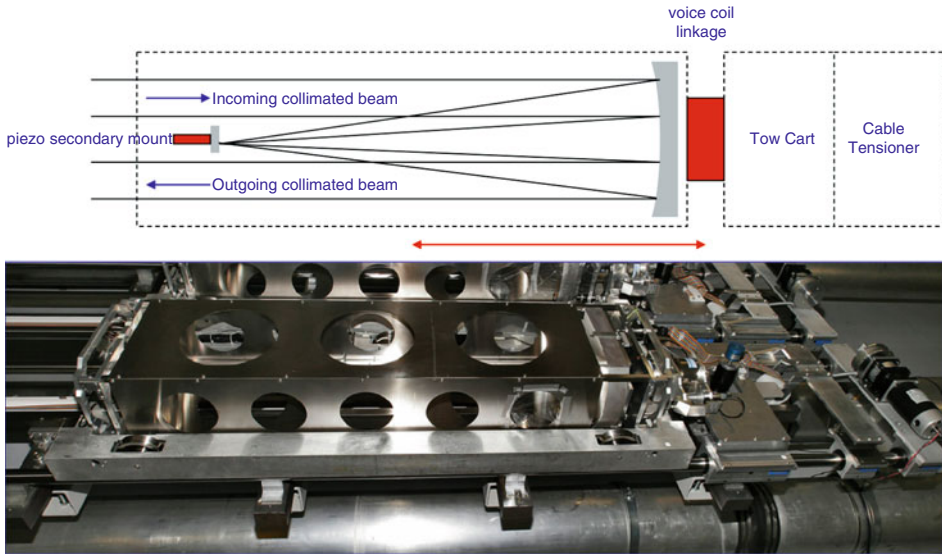
The total amount of path length correction required is a direct function of the size of the baselines and how far away from zenith the observations might take place. For example, a 100-m baseline going to 70° from zenith requires approximately 64 m of controllable delay. Furthermore, this delay must be controlled to a small fraction of the size of the observational wavelength. This calls for quite a large and complex optomechanical system capable of introducing paths many hundreds of meters long to a precision of a fraction of a wavelength, 10 nm or so, or about one part in 10^{10} .

In most cases, the OPLE system is broken up into two separate parts: one to introduce large but static delays and one that has a smaller but variable delay. The first can be implemented by using numerous fixed mirrors on movable stages that flip in and out of the beam, sometimes referred to as “POP” mirrors, and, in comparison with many of the optical systems in an interferometer, these are relatively simple systems. The second continuously variable part of the OPLE is more difficult but, in almost all cases, makes use of what is known as a *cat’s eye*. Because POPs introduce more reflections, in some cases they are not used at all, which will either reduce sky coverage, maximum baseline length, or imply the need for very long optical rails. A more detailed description of these types of optomechanical systems can be found in Colavita et al. (1991), Frederic (2000) and Fisher et al. (2010).

A cat’s eye is an optical system that returns the outgoing beam very closely parallel to the direction of the incoming beam. Cat’s eyes are very insensitive to the angle of arrival, that is, the cat’s eye can be rotated with respect to the incoming beam without changing the direction of the outgoing beam. They most often consist of a parabola and a small flat mirror as shown in  Fig. 6-7. These cat’s eye assemblies are mounted on small carts that move along optical rails and, because they are double pass, introduce an optical delay twice as large as the movement of the cart along the rail. These rails need to be quite long, extremely long in the case of interferometers that do not have POPs.

A final requirement for the OPLEs is some way of measuring the true optical delay. Without this it would be impossible to close the control servo loop. This is normally done by passing a metrology laser through a different axis of the cat’s eye and using that to measure the position of the cart. It is also sometimes useful to use “end-to-end” metrology where the full internal path all the way from the beam combiner out to the telescopes is measured. This is particularly important in astrometric applications where the position of objects on the sky needs to be measured with very high precision. If the positions of the telescopes are known very accurately and the position of the OPLEs required to find the fringes is recorded, the $OPD = B \sin \theta$ relationship can be inverted to solve for θ .

For even greater precision, more metrology systems are required in order to measure the locations of the telescopes themselves in all three dimensions (Hutter and Elias 2003). Finally, it is also possible to achieve extremely high-precision small-angle separation measurements by



■ Fig. 6-7

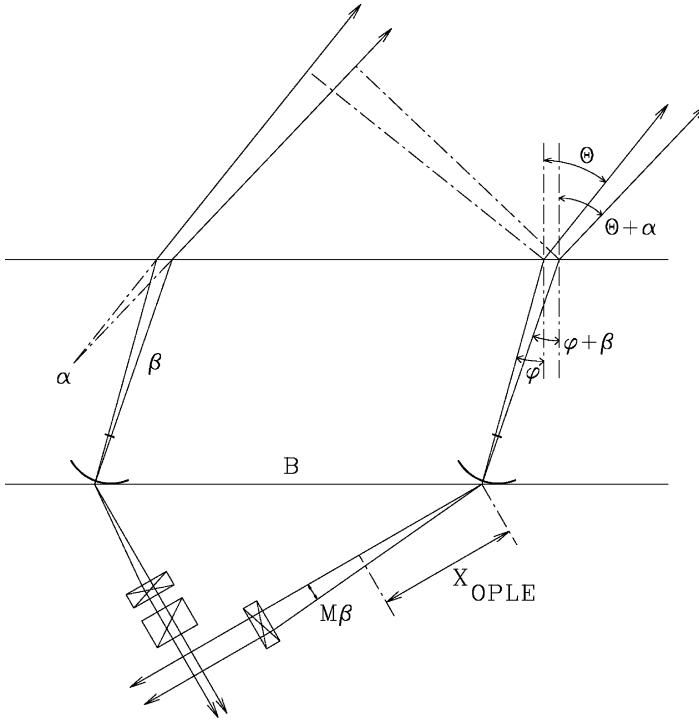
Bottom: A picture of the delay line carts of the CHARA Array Interferometer. **Above:** A schematic of the delay line cart. The collimated input beam enters one side, forms an image on the secondary mirror, and is then recollimated and exits the cart on the other side. A metrology laser goes through the same optical system but vertically to make it possible to measure the position of the cart and close the servo loop. The piezo can introduce path modulation at the submicron level, while the whole optical unit is mounted on an inverse pendulum driven by a voice coil, able to modulate the beam by a few centimeters. The cart itself is attached to a tow cart via a voice coil and the whole cart moves on a 50-m-long set of optical rails. The control system is a four-tiered nested servo with the piezo position updated at a rate 5,000 Hz, the pendulum at 200 Hz, the tow cart voice coil at 20 Hz, and the stepper motor at 2 Hz

breaking the incoming beam into two beams, one pointing at each of two objects, and having a delay line for each target. A measurement of the differential delay between the two carts is a direct measurement of the separation of the two objects in the direction of the projected baseline. See, for example, Woillez et al. (2010) and Delplancke (2008).

More exotic methods of controlling optical delay exist, for example, stretching an optical fiber, but they have yet to be put into practice in long-baseline interferometry, and, despite their complexity, the cat's eyes used in the current generation of interferometers more than meet their extremely tight design tolerances.

Differential Refraction and Dispersion – There are two types of effects caused by the atmosphere's refractive index that are of concern: angular refraction and longitudinal dispersion. ▶ *Figure 6-8* shows the basic geometry of the causes of refraction and dispersion. There is only enough space here for a brief discussion of these issues, but more complete discussions of refraction and dispersion can be found in Tango (1990) and ten Brummelaar (1995).

Differential angular refraction in an interferometer is very similar to that in a single-aperture telescope. Unless the object being observed is exactly at the zenith, the incoming light will enter the earth's atmosphere at an angle and, since the atmosphere has a refractive



■ Fig. 6-8

Diagram of the geometry used to calculate the differential path effects caused by the refractive index of air: The telescopes (curved lines) are mounted on the ground (lower horizontal line) and separated by the baseline distance B . The upper horizontal line represents the top of the atmosphere. The ARCs (one in each beam) are represented by the rectangles with crossed lines, and the LDC by the rectangle with the single diagonal line. φ is the zenith angle of the scientific target after refraction by the atmosphere. The dashed-dotted line represents the baselines of the pointing center and center of scientific interest

index different to that of the vacuum of space, this angle will change. Furthermore the refractive index of the air is wavelength dependent, so the change in direction will be different for each wavelength. This change in angle is given by

$$\Delta\theta = \left(\frac{1}{n_{\text{air}}(\nu)} - 1 \right) \tan(\theta),$$

where $n_{\text{air}}(\nu)$ is the wave number-dependant refractive index of air. In standard, imaging this results in a blurred or “banana”-shaped point spread function, but in an interferometer, the problem is slightly different. In any given band pass, each wavelength will have a slightly different angle of arrival, and so the geometric optical delay $B \sin \theta$ will be slightly different. Unfortunately, the delay line can only be in one position and so the different wavelengths will each have a different delay offset with only one, normally the one in the middle of the band, having a true zero delay. The result is a smearing of the fringes into a broader fringe envelope, with one side containing the “red” fringes and the other side containing the “blue” fringes.

This can be a serious problem for closure phase measurements when fringe overlap is not perfect, causing the combination of the red fringes with the blue fringes and thereby introducing noise into the phase signal.

A second, and more important, problem introduced by differential angular dispersion results from the fact that the light must be split up in order to divide it between the various detectors such as tip/tilt and the beam combiners. It is not unusual for the tip/tilt detector to operate at a very different wavelength to that of the beam combiners, and since by definition the tip/tilt will center the beam for its operational wavelength, the beams reaching the beam combiner will enter at a substantially different angle. Fortunately, the beams from all telescopes will be affected in the same way and will all enter the beam combiner at the same angle, albeit not the correct one. This can, however, result in a significant beam shear, causing vignetting and a loss of signal. This can be overcome by introducing offsets in the tip/tilt detector to find a compromise between signal loss in the tip/tilt system and signal loss in the beam combiner (Stomski et al. 2003). A similar problem results from the fact that the system can only have the correct delay for one position in the sky and for one wave length, the so-called *phase center*, and since the field of view and the band pass are both nonzero, this can also result in beam offsets and vignetting as shown in [▶ Fig. 6-8](#).

In both the single aperture and interferometer cases, there are two ways of dealing with differential refraction. First, the bandwidth of the spectral channels can be reduced as the smaller the bandwidth, the smaller the differential refraction. Secondly, atmospheric refraction can be corrected by using two optical wedges called *Risley Prisms* (Breckinridge et al. 1979) that can be rotated with respect to one another in a device called an atmospheric refraction corrector (ARC). In order to minimize the vignetting problem described above, the ARCs need to be installed at or near the telescopes. In most existing facilities, the beam size is so large and the required optical bandwidth is so broad that building ARCs is quite expensive and they are often not used.

Unlike single-aperture imaging systems, in an interferometer, not only changes in angle but changes in optical path lengths are a major concern. This introduces the other type of effect caused by the refractive index of the atmosphere: differential longitudinal dispersion caused by differences in air paths inside the instrument. If the air paths are not the same in each arm of the interferometer, each wavelength will “see” a different optical path. At best this creates the same increase of size in the fringe envelope, causing it to have a red side and a blue side, and at worst, it can destroy coherence altogether. Differential longitudinal dispersion is a much more serious problem than atmospheric refraction.

There are two potential sources of differential air paths in an interferometer. First of all, telescopes may be at different elevations and different distances from the central beam-combining facility. This can be ameliorated by sending the beams from each telescope through evacuated light pipes as discussed above. The vacuum need not be extreme as only enough of the air has to be removed to minimize the refractive index effects of the air, and a vacuum of a few Torr is sufficient for this. Secondly, the geometric delay caused by the object not being at zenith is in fact in the vacuum of space above the atmosphere, as shown in [▶ Fig. 6-8](#). This means that if the delay lines are in air, they are correcting a path length difference in a vacuum by one in the air and in this way can introduce very large amounts of differential dispersion. One way around this is to place the delay lines in a vacuum system, essentially very large light pipes, and this is the preferred method. This can, however, be expensive from the point of view of both construction and maintenance, as it means placing a very complex optomechanical system inside a vacuum, which will need to be broken any time an adjustment to the optics or other hardware is

needed. It also implies the need for vacuum-resistant motors and other materials. Finally, a long vacuum system can make servicing the carts very difficult should one get stuck in the middle of a vacuum delay line 100 m in length. The result is that in many cases the delay line is left in air.

Fortunately, like refraction, it is possible to correct for most of the different dispersion by using two glass wedges that move with respect to one another and thereby introduce more or less glass into each beam. This can be done in such a way as to introduce the same amount of dispersion using the glass in one beam that is introduced into the other beam by the air. In an interferometer, differential and not absolute effects are important, and so the total amount of dispersion in each beam is not important, only that each beam contains the same amount of dispersion.

This means changing (6.1) and introducing terms for the dispersion so the optical path length difference now becomes

$$\text{OPD} = B \sin \theta + \underbrace{n_{\text{glass}}(\nu)d}_{\text{LDC}} - \underbrace{n_{\text{air}}(\nu)x_{\text{OPLE}}}_{\text{OPLE}},$$

where d is the amount of glass in the LDC and x_{OPLE} is the air path within the delay line. The OPLE path is then defined as a displacement l from the position x_0 :

$$x_{\text{OPLE}} = l + x_0,$$

where x_0 is the position the delay needs to be to find fringes when there is no glass in the LDC and the instrument is observing at the phase center. Thus,

$$x_0 = \frac{1}{n_{\text{air}}(\nu_{\text{pc}})} B \sin \theta,$$

where ν_{pc} is the wave number that defines the phase center and this results in

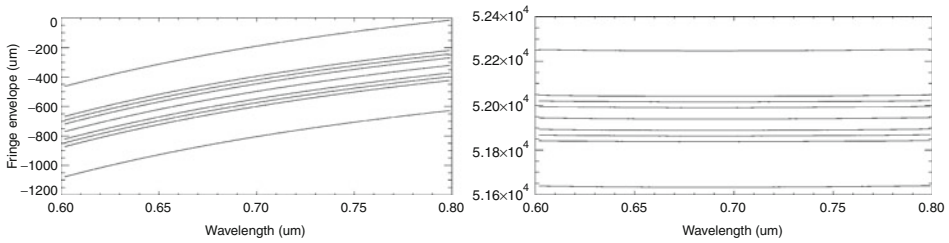
$$\Delta \text{OPD} = B \sin \theta \left(1 - \frac{n_{\text{air}}(\nu)}{n_{\text{air}}(\nu_{\text{pc}})} \right) + n_{\text{glass}}(\nu)d - n_{\text{air}}(\nu)l,$$

which can now be substituted into (6.1). Since the refractive indices of air and the glass used in the LDC are functions of ν , the integral in (6.1) is now nontrivial. The solution is to use a first-order Taylor expansion and assume that within the limited size of the band pass, the refractive indices of both air and of glass are approximately linear and that

$$n_{\text{air}}(\nu) \approx 1 - (\nu - \nu_0) \frac{dn_{\text{air}}}{d\nu}(\nu_0) + \mathcal{O}[(\nu - \nu_0)^2],$$

with a similar expression for $n_{\text{glass}}(\nu)$. This makes the integral in (6.1), while still messy, much easier to perform, and the result is a new fringe equation in the same form as (6.2) with a modified fringe phase, which, since closure phases are being used, can, to first order, be ignored. This leads to a much more complex expression for the optical path length difference in the sinc or envelope part of the fringe equation:

$$\begin{aligned} \text{OPD} = & B \sin \theta \left(1 - \frac{n_{\text{air}}(\nu_0)}{n_{\text{air}}(\nu_{\text{pc}})} - \nu_0 \frac{dn_{\text{air}}}{d\nu}(\nu_0) \right) + \left(\nu_0 \frac{dn_{\text{glass}}}{d\nu}(\nu_0) + n_{\text{glass}}(\nu_0) \right) d \\ & - \left(\nu_0 \frac{dn_{\text{air}}}{d\nu}(\nu_0) + n_{\text{air}}(\nu_0) \right) l. \end{aligned}$$



■ Fig. 6-9

Left: The fringe envelope for a baseline of 100 m and a zenith angle of 50° without dispersion correction. The *central line* represents the center of the envelope. The succeeding *pairs of lines* represent, respectively, the 95% point, the 90% point, and the 85% point. The *outermost lines* represent the first zero. The Y axis shows the delay with respect to zero point x_0 . **Right:** The same fringe envelope but with the correct amount of glass, in this case BK7, for dispersion correction

Thus by knowing the geometry of the array, the position of the target star, the amount of air and vacuum paths within the interferometer, the functional form of the refractive index in air, and the glass used in the LDC, the amount of glass required at any given time to balance out the effects of differential dispersion can be calculated. An example of this is given in [Fig. 6-9](#).

The left plot of [Fig. 6-9](#) shows that across an optical band of $0.2 \mu\text{m}$ and a baseline of 100 m, the fringes in the center of the band have been shifted by $0.5 \mu\text{m}$ from x_0 , and the envelope center has moved by $0.5 \mu\text{m}$ within the band. The first of these can be corrected simply by moving the delay line $0.5 \mu\text{m}$, but the second will completely smear out the fringes and no signal will be seen at all. When the correct amount of glass, in this example BK7, is inserted, the change in delay across the band is reduced with the result of removing fringe smearing. Note, however, that the whole fringe packet has now been moved some 5 cm away from the default position, and this implies moving the delay line by that amount also.

Astrometric Model – In order to correctly position the delay lines and find fringes, an accurate mathematical model of the entire interferometer, known as the *astrometric model*, must be created that includes the array geometry, telescope pointing models, air and vacuum paths, POPs, ARCs, LDCs, and any path length changes introduced by beam-switching optics within the beam-combining laboratory. As shown in the previous section, many of these systems are coupled. For example, if the delay line is in air, changing its position will introduce differential dispersion which means the LDC must be moved to correct for this. Moving the LDC will then introduce changes in optical delay and require a change in the delay line position, thereby once again changing the differential air paths and differential dispersion. This process converges, but as the earth rotates and the position of the star in the sky changes, all of these optical systems are continually moving. In the case of the telescopes and delay lines, this requires constant motion in order to track the star and the fringes, with frequent new predictions of the position and velocity of motion of the telescope axes and the delay line. In the case of the telescope axes, this is necessary a few times per second, while in the case of the delay line this need only be done every few seconds. The ARCs and LDCs do not need to be in constant motion and can be moved to a new position in between data integration cycles. This helps circumvent the coupling between the delay line and the LDC.

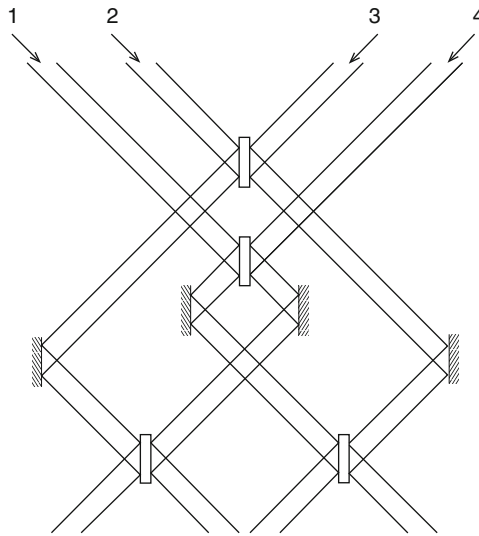
It is rare that this model can correctly place the delay lines such that fringes appear immediately – there are just too many variables – and there will almost always be a need to search for the fringes. For example, a one arcsecond error in the alignment of a mirror that sends a beam down 200 m of light pipe will result in a beam shift of 1 mm, which if reflected on a 45° mirror, will introduce a path length change also of about 1 mm, representing an error some 30 times the size of the typical envelope width. Similar errors can be introduced by misalignments in the mirrors within the telescopes and these will change with position on the sky. Nevertheless, the better the astrometric model is, the less time will be spent searching for fringes. To do this, it is necessary to know the positions of the telescopes and the internal path lengths to very high precision. To know a 100-m baseline to a tolerance of 1 mm represents one part in 10^5 , and no ordinary surveying technique will yield this kind of accuracy. The solution is to find fringes on many targets across the sky, record the delay line and LDC positions required, and then solve for the telescope positions and internal paths. Here, the position of the telescope is defined as the intersection of the rotational axes. This is known as performing a *baseline solution*, and it is the equivalent of a pointing solution for a single telescope, which of course must also be done for each of the telescopes in the array. In many cases, the data required for the baseline solution can be collected in the normal course of an observing season.

Control System – Interferometers are made up of numerous subsystems and servo systems, many of which require active real-time control. They are also spread out over large areas. This means that the control system for an interferometer needs to be a highly distributed multiple-CPU real-time network of computers. Furthermore, many of the servo systems are coupled and must interact, for example, the LDC and OPLE systems discussed above, and the tip/tilt servo and telescope pointing systems. Another very important requirement is that all of the computers in the control system must agree on the time, and that this time is very accurate, at least better than within 1 mS of the true Universal Time. It is also necessary to have some way to coordinate the various subsystem control computers in order to correctly sequence the actions necessary to collect data. For example, all the telescope control systems must acquire the same target, the tip/tilt system needs to be locked, and the delay lines, LDCs, and ARCs must be in the correct position, all before it is possible to begin searching for the fringe signal.

6 Beam Combination

Given that all the telescopes are pointing at the same object, all the mirrors in the system are correctly aligned, and all the delay lines, ARCs, and LDCS are in the correct positions and tracking, there will be a collection of beams reaching the beam-combining laboratory that are all in phase. What remains is to find a way to combine these beams in an efficient manner. This is similar to having a single-aperture telescope well aligned and pointing at the sky – without scientific instruments at the back end no science can be done. The beam combiner is the back end instrument of an interferometer and, as with a single-aperture telescope employing multiple focal-plane instruments, there will often be multiple beam combiners, each optimized for a particular kind of observation and science goal, employing different wavebands from the optical to the infrared, and with various spectral resolutions, numbers of beams, and sensitivities.

There are so many different approaches to designing a beam combiner, it is not possible to deal with them all here, but the most important elements of the design choices will be discussed.



■ Fig. 6-10

A four-beam aperture-plane beam combiner based on the COAST design (Baldwin et al. 1998). The four beams are mixed together and are sampled in each of the four output beams. The output beams are all sampled by a single-pixel detector, or by multiple pixels if spectrally dispersed

Aperture-Plane Combination – In this method of beam combination, the beams are combined using beam splitters in something that is sometimes erroneously called a Michelson interferometer despite the fact that the first interferometer to measure a stellar diameter was built by Michelson and Pease (1921) and did not use this method. ● Figure 6-10 shows a simple bulk optic four-way aperture-plane beam combiner where two sets of pairs of beams are combined on beam splitters, and then these two pairs of beams are combined in a second set of beam splitters. This kind of layout can be expanded to more beams but is restricted to powers of two, and, of course, many other similar approaches exist for a range of numbers of beams.

In this beam combiner, there are four beams in and four beams out, and if there is no spectral resolution, only four single-pixel detectors are required for each spectral channel or only four pixels on a multipixel detector. This has the advantage of concentrating as much light as possible into the smallest number of pixels, and this will give the highest possible signal-to-noise ratio in each pixel and, therefore, the best sensitivity. It is for this reason that sensitivity is the primary advantage of an aperture-plane beam combiner.

The four beams will produce four different pairs and four sets of fringes, representing four different baselines, and all four of these are present in each of output pixel data streams. It is necessary to analyze each fringe signal separately, so these four sets of fringes must be encoded in some way and this is normally done by modulating the optical path length in each telescope, sometimes by using the delay lines and sometimes with the use of a mirror mounted on a piezoelectric stack called a *dither mirror*. In this way, a time series of samples of each output will have four superimposed fringe packets each one very much like those shown at the bottom of ● Fig. 6-2. This technique is known as *temporal fringe encoding*. This path length modulation must be done so that each set of fringes has a unique fringe frequency so that band-pass filters

centered on each of these frequencies can then be used to separate the signals of the four baselines. In practice, the Fourier transform of each fringe scan is used to extract a complex fringe amplitude at each of the four frequencies that gives a visibility amplitude and phase for each baseline.

For example, if beam 1 is not modulated and the frequencies of the other beams are f_2 , $f_3 = 2f_2$, and $f_4 = 2f_2$, the frequencies for the four baselines will be

$$\begin{aligned} f_{12} &= f_2, \\ f_{23} &= f_2 + f_3 = 3f_2, \\ f_{34} &= f_3 + f_4 = 4f_2, \\ f_{14} &= f_4 = 2f_2. \end{aligned}$$

So all that is left is to choose the frequency for beam 2. There are other constraints that must be taken into account. Each of these scans must take the same amount of time, and so the stroke of beams 3 and 4 will need to be half that of beam 2, and the length of this stroke must be larger than the fringe envelope, as well as large enough to take into account any changes in fringe location introduced by the atmosphere. Furthermore, in order for these frequencies to add as set out above, the motion of the stroke of beam 3 must be in the opposite direction to that of beams 2 and 4. The minimum fringe frequency must also be fast enough to ensure that the fringes between beams 1 and 2 are scanned in less than the coherence time of the atmosphere τ_0 . There will also be a maximum possible sample rate set either by the detector itself or by the magnitude of the star being observed, and because fringe phases will be required in order to calculate phase closure, the fringes between beams 2 and 3 must have at least three samples per fringe.

Finally, atmospheric seeing will modulate the fringe frequency in each channel, and so the peaks in the power spectrum can be quite broad. If these peaks are not separated by more than the atmospheric broadening, there will be cross talk between the baselines, and this will reduce both the quality of amplitude calibration and the precision of the closure phase measurements. All of these things need to be balanced and will change as the atmospheric seeing changes.

Normally, the fastest sample rate t that still has enough photons per sample is chosen, and this then determines the rate of motion of the strokes in beams 3 and 4 so that $f_3 = f_4 = 3/t$. The velocities of motion of the dither mirrors are then set to

$$\begin{aligned} v_2 &= \frac{f_2 \lambda_0}{2}, \\ v_3 &= -2v_2, \\ v_4 &= 2v_2, \end{aligned}$$

where λ_0 is the central wavelength of the current band pass and the factor of 2 is there because these dither mirrors are normally double pass. A normal data set would consist of a few hundred fringe scans followed by an equal number of scans consisting of some scans with the shutters on all beams closed, and four sets of scans each with only one beam reaching the beam combiner. These so-called *shutter sequences* are required in order to measure the background counts and to measure the value of the transfer function T_{ij} for each baseline. They also yield the mean photon count in each beam which is used to compensate for unequal beam intensities and are useful for calculating the noise in the fringe power spectra.

A visibility estimate can then be obtained by integrating the power in the fringe power spectra. The power spectrum of the fringe equation (6.2) is

$$\text{PS}[I_{ij}(v_0, \Delta v)] = T_{ij}^2 \left(\frac{V}{\Delta v g_{ij}} \right)^2 \left[\Pi \left(\frac{V - v_0 g_{ij}}{\Delta v g_{ij}} \right) + \Pi \left(\frac{V + v_0 g_{ij}}{\Delta v g_{ij}} \right) \right],$$

where g_{ij} is the group velocity of the baseline set by the speed of the dither mirrors and

$$\Pi(x) = \begin{cases} 0 & \text{for } x < -0.5 \text{ and } x > 0.5, \\ 0.5 & \text{for } x = -0.5 \text{ and } x = 0.5, \\ 1.0 & \text{for } -0.5 < x < 0.5 \end{cases}$$

is the top-hat function. Since the fringe signal is real, this can be simplified to

$$\text{PS}[I_{ij}(v_0, \Delta v)] = T_{ij}^2 \left(\frac{V}{\Delta v g_{ij}} \right)^2 \Pi \left(\frac{V - v_0 g_{ij}}{\Delta v g_{ij}} \right),$$

and the total integrated power is then

$$\int \text{PS}[I_{ij}(v_0, \Delta v)] df = T_{ij}^2 \frac{V^2}{\Delta v g_{ij}}.$$

Since the transfer function has been measured and the group velocity is known from the dither mirror velocities, this is a good estimate for the visibility amplitude. Note that if there is more than one baseline the limits of this integral must be carefully chosen in order to include all the power for that baseline without including any other.

Since it is extremely unlikely that τ_0 is larger than the time it takes to sweep through all four fringe packets in order to measure closure phase, it is necessary to break each scan up into smaller segments with each segment equal to the time it takes to scan through a single fringe of the slowest beam 1 and 2 baseline, which will be $1/f_2$. This segment will also contain two fringes of the beams 1 and 4 baseline, three of the beams 2 and 3 baseline, and four of the beam 3 and 4 baseline. A Fourier transform, which in practice will in fact be a discrete Fourier transform, will contain in the first four bins a measurement of the complex visibility for each baseline including both amplitude and phase. It is from these phases that closure phases can be formed, though it is extremely important to very carefully consider the signs of these phases.

For example, consider the closure phase signal for the first three beams, that is, Ψ_{123} . The phase of beams 1 and 2 is positive by definition, and we wish to add the phase of beams 2 and 3. Note, however, that the dither mirror for beam 3 is moving in the opposite direction and so this phase has a negative sign. Furthermore, a closure-phase triangle must be closed and so it is the phase of beams 3 and 1 that is required, and so it too has a negative sign. This is more easily expressed as a triple product of the three complex visibilities, and using the fact that a sign change of the phase is the same as taking the conjugate of a complex number, we find that

$$V_{123}(v) = V_{12}(v) \times \text{conj}[V_{23}(v)] \times V_{13}(v).$$

This can be calculated for all the fringe segments within the area where all three fringes overlap for all scans to find a mean triple product that will contain an amplitude weighted mean of the closure phase. A more comprehensive description of the analysis of fringe data can be found in Monnier (2011).

Finally, it is also possible to obtain an estimate of the atmospheric coherence time from these data. As atmospheric seeing changes, it will introduce variations in the group velocity of each


fringe scan. This is what causes the broadening of the peaks in the power spectrum mentioned above. Thus, the peak of the power spectrum for each scan will be in a slightly different place. If the peak in a given scan is found at frequency f_{\max} , the group velocity will be given by


$$g = \frac{f_{\max}}{\nu_0}.$$

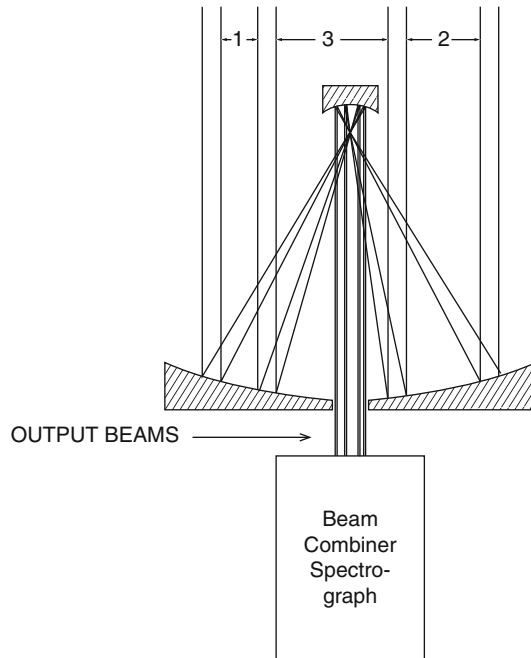
The mean of this group velocity should be the same as the group velocity defined by the motion of the dither mirrors for that baseline, while the standard deviation of the group velocity σ_g is related to the coherence time of the atmosphere via

$$\tau_0 \approx \frac{\lambda_0}{2\pi\sigma_g}.$$

To summarize, an aperture-plane beam combiner uses bulk optics, such as beam splitters and mirrors, to bring the light from all telescopes together. This has the advantage that it uses the minimum number of pixels and therefore the maximum amount of light and, in the case of detectors with readout noise, signal-to-noise ratio in each pixel, giving it the best possible sensitivity. The disadvantage is that it requires temporal fringe encoding, that is, changing the path lengths of each beam in time, which imposes limits on the sample times in the detector and involves precise and fast motions of the delay line, or a separate dither mirror, with the added danger of cross talk between the baselines. Furthermore, if the delay lines are used, the path length modulations will be present in all parts of the beam-combining optics including all other beam combiners. Finally, it is not possible to modulate the paths fast enough to freeze the atmosphere through the area of fringe overlap without imposing impossible constraints on sample time and sensitivity, and so, residual atmospheric noise will always be present in the amplitude and closure phase measurements.

Image-Plane Combination – An alternative beam combination method is to combine the beams in the image plane by using lenses or other powered optics as in the example image-plane combiner shown in  Fig. 6-11. In this example, the box marked “Beam Combiner Spectrograph” contains the power optic that forms the actual fringes.

This will only work if the image is close to being diffraction limited, that is, the telescope diameters must be small or must have adaptive optics. An alternative is to use spatial filtering which will be covered in the next section. A second requirement is that the telescopes do not resolve the object, but there would be no point in building an interferometer if this were the case. The result is that the image plane will contain the point spread function of the optical system across which there will be one set of fringes for each pair of beams. As with Young’s double slit experiment discussed above, the spatial frequency of these fringes will depend only upon the spacing of the beams as they enter the powered optic. Note that in  Fig. 6-11, the beams are arranged so that the spacing between any two beams is unique, and this results in the fringe frequency of each fringe pattern being unique. This is known as *spatial fringe encoding*, and as for aperture-plane interferometry, this is required so that it is possible to extract the fringe information for each baseline without cross talk. Also like aperture-plane interferometry, it is possible to disperse the light and obtain spectral information, but in this case it is first necessary to compress the image perpendicular to the fringe axis before adding dispersion in order to avoid spectral channel overlap. A method for doing this will be discussed in the next section. A Fourier transform of the image plane will contain several peaks, one for each baseline, and these data can be analyzed in a manner analogous to that of aperture-plane data.



■ Fig. 6-11

A four-beam image-plane beam combiner. The four beams enter from above with nonredundant spacing and are reduced in size by a pair of parabolas. The four output beams are in the same nonredundant pattern but are now small enough to pass through a lens, forming an image with fringes inside the *box* at the *bottom*. This requires a multipixel detector

The principal advantages of image-plane interferometry are twofold. First, it normally involves fewer optical components than aperture-plane systems. Second, all fringe amplitudes and phases are sampled simultaneously, and so unlike aperture-plane systems, where the sample time must increase as you add baselines and thereby force an increase in the maximum fringe frequency, the sample rate of an image-plane system is independent of the number of baselines. So long as the sample time of the detector is less than τ_0 , there is much less atmospheric smearing of the fringes and less noise in the closure phases. An image-plane beam combiner is relatively easy to expand to very large numbers of beams while an aperture-plane combiner is not.

There are, of course, disadvantages, and the most significant one of these is that this method requires many pixels in order to properly sample the image plane. In the example system shown in [Fig. 6-11](#), the maximum spatial frequency is six times that of the minimum. Like aperture-plane systems, we require better than Nyquist sampling, so assuming a minimum number of three samples for the highest spatial frequency, it turns out we need 18 pixels for each spectral channel as opposed to the four in the aperture-plane beam combiner shown in [Fig. 6-10](#). Thus, each pixel will have 4.5 times less light and therefore a much lower signal-to-noise ratio.

Field of View – The field of view of an aperture-plane interferometer is a function of the baseline and the fringe envelope size. The telescopes, LDC, ARC, and delay lines are nominally positioned such that the center of the image is the phase center defined above. At an angle $d\theta$

away from the phase center, the geometric delay will be $B \sin(\theta + d\theta)$ and, assuming this angle is small, this means there is a delay error of approximately $B \times d\theta$. If this reaches the first null in the fringe envelope, no fringes will be seen, so the edge of the field of view will be reached when

$$B \times d\theta < \frac{1}{\Delta\nu},$$

and thus the full field of view of the instrument is

$$\text{FOV} = \frac{2}{B\Delta\nu} = \frac{2\lambda^2}{B\Delta\lambda}.$$

This can be a very small angle. For example, for a 100-m baseline in the near-infrared K band, the field of view is 0.13 microradians or 26 milliarcseconds. Furthermore, the field of view will be determined by the maximum baseline currently being used.


The field of view of an image-plane interferometer is subject to the same path length restrictions as the aperture-plane interferometer discussed above, but there is another restraint set by the aperture size of the telescopes. This arises if the point spread functions do not overlap in the image plane so that no fringes can be formed and sets an upper limit for the field of view independent of the baseline size. So, for example, for a circular aperture, the point spread function is an Airy disk whose size is inversely proportional to the telescope aperture size given by $2.4\lambda/D$. For a 1-m diameter telescope in the near-infrared K band, this will be about one arcsecond. At shorter wavelengths, this upper limit really comes into play, for example, at $0.5 \mu\text{m}$, it is only 0.25 arcseconds.


Using a technique known as *Fizeau interferometry* can substantially increase the field of view of an image-plane beam combiner. In a Fizeau interferometer, the arrangement of the apertures that enter the beam combiner are not necessarily nonredundant but are in the same pattern as the projection of the telescopes on the sky. This brings the idea of an interferometer as being a large lens to its logical conclusion. Such an arrangement will only be limited by the size of the telescopes and not by the internal paths inside the beam combiner. One serious difficulty of Fizeau interferometry is that the projected arrangement of the telescopes on the sky is continually changing, and this means that the output apertures of the beam combiner must also continually change. This is extremely difficult as this motion must also be stable to within a small fraction of the central wavelength. Another is that the aperture is very dilute unless there are a very large number of telescopes. At the time of writing, no one has succeeded in doing this.

Optical Fibers and Spatial Filtering – As [▶ Figs. 6-10](#) and [▶ 6-11](#) show, beam combiners are very complex optical systems and they require many optical surfaces and large areas of optical table space in a complexity that goes up geometrically with the number of input beams. They are also restricted by the wavefront quality of the incoming light, for without an adaptive optics system, it is not possible to use an aperture size of more than about $3r_0$ without losing all coherence due to atmospherically induced distortions in the wavefronts. Another problem when constructing beam combiners is how to arrange the incoming beams in the correct pattern while still preserving all the internal path lengths. For example, the beam combiner in [▶ Fig. 6-11](#) shows the four beams entering the system in a nonredundant spacing, but it does not show the optics required to get them into that pattern without introducing differential paths. One elegant way around all of these difficulties is to use single-mode optical fibers as beam transport mechanisms instead of bulk optics.

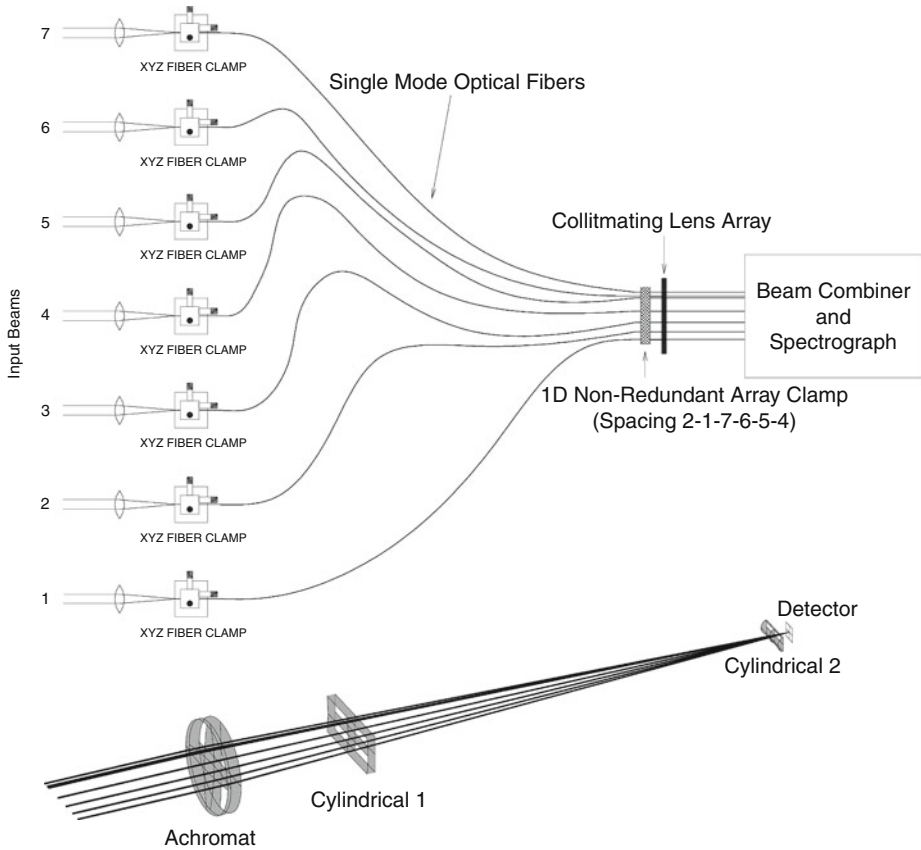
Single-mode fibers have the advantage of being able to be moved around to form any arrangement of output aperture spacing desired without the need to add more optical components. They are also very small and so very complex optical systems can be built without the need for vast areas of optical table space. The most important advantage of single-mode fibers is, however, the fact that they are a coherent transport mechanism, that is, while it is not possible to get all of the light from a telescope into a fiber, whatever light is coupled into a single-mode fiber will be coherent and completely free of wavefront distortions. This is known as *spatial filtering* and is the equivalent of using a lens and a pinhole to clean up the output of a laser on an optical bench.

As previously mentioned, a lens is a Fourier transform device, so the light focused onto a fiber, or a pinhole, is the Fourier transform of the incoming wavefront. A perfectly flat wavefront has no high spatial frequency power in its transform if it is infinite in extent. A finite-sized beam will have edges, which do contain high spatial frequency information, and it is this that results in the formation of the rings in an Airy disk. A distorted wavefront will have more high spatial frequency power and a broader point spread function, but the pinhole or fiber in a spatial filter will only allow the very-low-spatial-frequency light through. The result is that the final wavefront is of very high quality, albeit with a great deal of the light missing, the light that would not form fringes anyway. This is, in effect, trading wavefront quality for throughput and turning atmospheric distortions into intensity fluctuations.

Consider the image-plane beam combiner for seven beams shown in  Fig. 6-12. The fibers are used to bring the beams into a one-dimensional nonredundant pattern, which is compressed in the orthogonal vertical axis by cylindrical lenses which can then be spectrally dispersed. In this way, this system can provide both high spatial and high spectral information. Note that the nonredundant pattern has been chosen to minimize the total separation of the first and last beams, which in this case is 25 times the minimum spacing. If this were to be done using bulk optics, this minimum spacing would need to be at least twice the beam diameter so that the outside edges of each beam are at least one beam diameter apart. If the beams are one inch in diameter, the total width of this output array would be 50 in., while in the case of the optical fibers whose outside diameter is only a few tens of microns, this can be as small as a few millimeters.

The beam combiner shown in  Fig. 6-12 is the fiber equivalent of an image-plane system, but there are also fiber equivalents of aperture-plane devices. The communications industry has made enormous investments in fiber technology, and there are fiber-based equivalents of almost every kind of optical device from beam splitters through to dispersive elements for spectral resolution. Once the light is inside the fibers, the entire beam-combining operation can be performed inside fiber optics. The first beam combiner to make use of optical fibers was the Fiber Linked Unit for Optical Recombination, or FLUOR instrument (Coudé du Foresto et al. 1998), first installed at the now decommissioned IOTA interferometer (Dyck et al. 1995) and now still in operation at the CHARA Array. FLUOR still achieves the highest precision-calibrated visibility amplitude measurements of all optical or infrared beam combiners to date.

A more recent technology now being applied to optical interferometry is *integrated optics*. Integrated optics is the equivalent of integrated circuits in the electronics industry where complex fiber-optic based systems are etched into a substrate. It is now possible to make full beam combiner systems on a piece of glass the size of a match. These integrated optics systems have recently been used on the sky (Benisty et al. 2009) and have enormous potential for the miniaturization and simplification of beam combiners. For the time being, however, they are very



■ Fig. 6-12

A seven-way beam combiner that uses single-mode fibers matched in length to arrange the beams into a linear nonredundant pattern. The *box on the top right* contains the optics shown on the *bottom* of the figure. The *cylindrical lenses* compress the final image into a small vertical space while the fringes appear across the horizontal axis. This can then be spectrally dispersed in the vertical axis, providing both spatial and spectral resolution. The Michigan Infrared Combiner (Che et al. 2010) is a version of this type of beam combiner

expensive, and those building integrated optics beam combiners have often had to rely on “piggy backing” their chips’ manufacture in the unused parts of substrates in large commercial runs. It is likely, however, that this will change in time, and integrated optics will become a common component in optical interferometers.

As might be expected, there is a price to pay for the use of optical fibers, and that is the difficulty of coupling the light from telescopes that can be many meters in diameter into single-mode optical fibers that are a few microns across. The beam shape of an optical fiber is Gaussian, while that of a circular aperture is an Airy disk, so even in a perfect world it is not possible to have perfect coupling into a fiber. The best coupling is when the Airy disk and Gaussian shapes are matched as well as possible and this yields a coupling efficiency never better than 78%. In the real world, this is rarely achieved, and, when atmospheric distortions of the

wavefront are taken into account, it is rare to get even close to this maximum so that coupling efficiencies of only a few percent are not uncommon. Adding to this, as the atmospheric seeing changes in a chaotic way, the coupling efficiency will also fluctuate chaotically with the result that there will be large fluctuations in light intensity, some of which will be at the fringe frequency. This can add a large amount of noise to the fringe signal. The problem of the overall amount of coupling can only be overcome by improvements in the wavefront quality with the use of adaptive optics while the increased intensity fluctuations must be measured and calibrated out or else they can overwhelm the fringe signal (Coudé du Foresto et al. 1997). The result is that while single-mode fiber-based beam combiners produce the highest quality data, they are very limited in sensitivity. If higher sensitivity is required, it is best to use bulk optics instead.

There are other difficulties associated with the use of single-mode fibers. While fibers do an excellent job of removing wavefront aberrations, they can do nothing about path length modulations, or *piston error*, introduced by the atmosphere. The fibers must also be polarization preserving or otherwise the polarizations will mix and the fringe coherence will be lost. Furthermore, fibers are not perfect light conductors and can lose several dB of intensity over very short distances. Another difficulty is that the glass from which fibers are made is a highly dispersive medium, and, in order to avoid introducing large amounts of differential dispersion, the fibers used in a beam combiner must be matched in length to very tight tolerances. Finally, although there are numerous fibers on the market produced for the communications industry, the development of fiber technology has been concentrated on very specific wavelengths, and procuring fibers for other wavebands of astrophysical interest can be difficult and exceedingly expensive.

Spatial filters can also be used without fibers. For example, the PAVO beam combiner (Ireland et al. 2008) is a kind of hybrid beam combiner where the incoming beams are set out in a nonredundant linear pattern in the same way as is done in an image-plane beam combiner, but the telescope pupils are reimaged and the fringes are formed in the aperture plane. PAVO includes spatial filtering by passing the light through very small square apertures. Another way of performing some spatial filtering has been used in the CLassic Interferometry with Multiple Baselines or CLIMB beam combiner (Sturmann et al. 2010). CLIMB is a fairly standard three-beam aperture-plane beam combiner that includes some spatial filtering by taking advantage of the fact that the camera pixels are of finite size. Spatial filtering is included in CLIMB by carefully ensuring that the spot size of the beams on the camera is slightly larger than the pixels on which they land.

Polarimetry – Part of the fundamental design constraints of the optics in an interferometer has been to preserve optical polarization in order not to mix different polarization states in the beam combiner. This has the added bonus that it is possible to add a few small optical elements in the beam combiner, such as a *Wollaston prism*, that can separate the polarization states on the detector, allowing the beam combiner to also serve as a polarimeter. In this way, images can be created that show the morphology of the different polarization states. Interferometric polarimetry (Elias 2001, 2004) is a relatively new part of interferometry, but almost all modern beam combiners recently or now being built include a polarimetry capability.

Fringe Tracking – A beam combiner has two important roles in an interferometer. Not only does it need to record the fringe data for later analysis but it must also *fringe track*, that is, locate and lock on to these fringes by sending offsets to the delay lines. The astrometric model is never perfect, and the atmosphere is constantly moving the fringes around so this is a very important part of an interferometer.

In the simplest form, the fringe-tracking signal is intended only to keep the fringe envelope within the scanning range of the dither mirrors in an aperture-plane beam combiner or in the correct position in the spectrally dispersed power spectrum of an image-plane beam combiner. This is called *group delay tracking*. Consider a single scan of a fringe in a single baseline of an aperture-plane beam combiner. This scan will contain a fringe pattern like that defined in (6.2), but the fringe envelope will not necessarily be centered in the scan. It is possible to demodulate this fringe data and thereby locate the center of the fringe envelope in the following manner. First, the data are band-pass-filtered at the known fringe frequency. Then, by the use of a modified Hilbert transform (Bracewell 2000), it is possible to demodulate this signal and obtain the envelope function. This is done by setting all the negative frequencies of the Fourier transform of the scan to zero and then calculating the modulus of the inverse transform. The distance of the center of this envelope function to the center of the scan is the current path length error and can be used as an offset for the delay lines.

Numerous other methods for group delay tracking exist, but they all have the same function of providing a measurement of the piston error, that is, the distance of the center of the fringe packet from where it should be. Group delay tracking is a very good technique for ensuring that fringes stay within the range of the dither mirrors, or roughly in the correct position of the image-plane beam combiner, but the group delay signal can only be provided at the same rate at which the scan data arrives. This is rarely even close to the coherence time of the atmosphere. The result of this is that group delay tracking does not enable longer integration times in the beam combiner, and it is still necessary to sample the data at rates faster than $1/\tau_0$.

An alternative method is known as *phase locking* in which the fringe tracker holds the fringes to a tolerance of much less than a single wave. For example, if the scan length of the dither mirror is reduced so that it is exactly one wavelength long, and this is done in such a way that there are exactly four samples across the scan, it is possible to extract both the fringe amplitude and the fringe phase from these data. It is also not too difficult to do this in a time less than τ_0 and in this way provide an error signal for the delay lines that makes it possible to lock onto a single fringe within the fringe envelope.

If the four parts, labeled A, B, C, and D, of the scan are in discrete steps, it is relatively easy to extract the fringe phase, which will be given by

$$\psi_{ij} = \tan^{-1} \left(\frac{B - D}{A - C} \right).$$

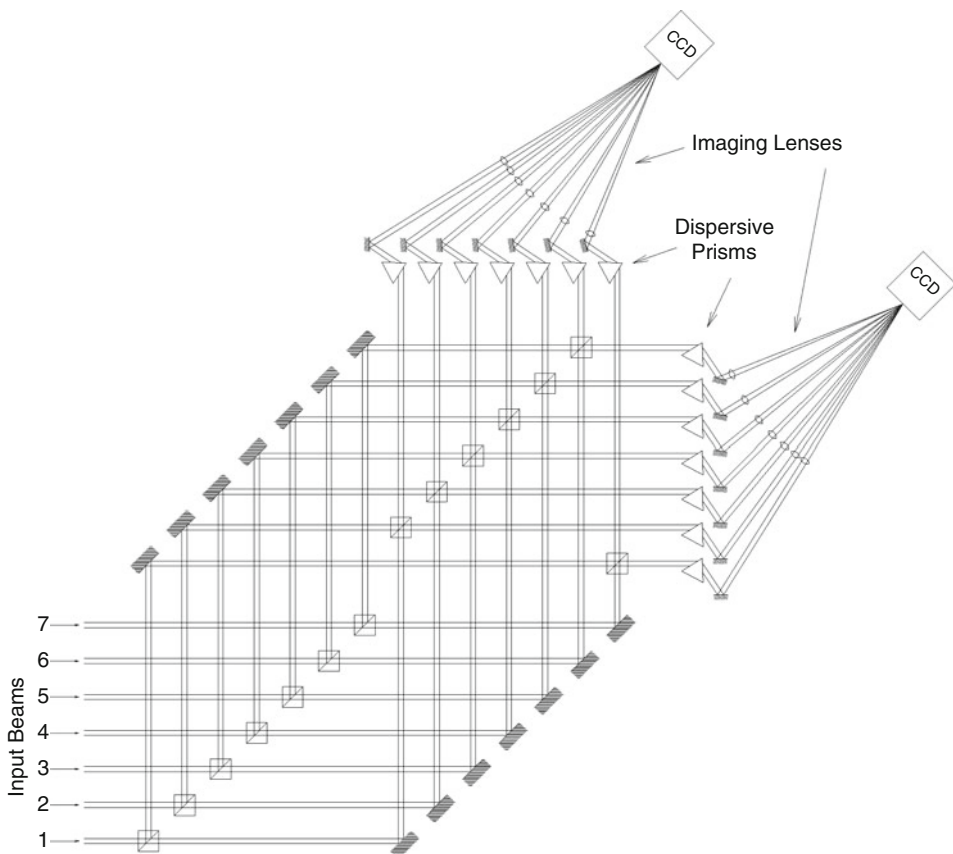
When this is multiplied by the central wavelength of the band pass, it gives a delay offset, which can then be used as an error signal for the delay lines. In practice, the scan is continuous and not discrete, and this adds a number of biases and a little more complexity to the analysis but with a very similar phase estimator (Colavita et al. 1999). These four data points will also provide a visibility amplitude estimate (Colavita 1999).

Phase tracking can provide the necessary error signal to lock the delay of the interferometer to within a fraction of a wavelength inside a single fringe, but it gives no information about which fringe inside the fringe envelope this is. In practice, both group delay tracking and phase locking are required to lock the delay lines onto the central fringe of the fringe packet.

Separating Fringe Tracking from Science – In the first generation of interferometers, a single beam combiner was required to both provide the fringe tracking information and also collect the data needed for the scientific program. This worked quite well, but it does impose extra constraints on the design of the system which then need to be optimized for both collecting

imaging data and for fringe tracking. These constraints are sometimes contradictory. More recently, these functions have been separated into two different optical systems, one for fringe tracking and one for the collection of the scientific data (ten Brummelaar 1994).

One of the main differences between a fringe tracking beam combiner and an imaging beam combiner is that a fringe tracker does not need to measure all fringes for every possible combination of telescopes, it merely needs to sample a subset of baselines that include each of the telescopes in the array at least once. If these baselines are correctly phased, then all possible baselines will be phased and the imaging system can then collect data on all baselines. This means that for N telescopes, the minimum number of baselines required for fringe tracking is $N-1$, although it is more common for symmetry reasons to form N baselines. An example layout for a seven-beam fringe tracker is given in [Fig. 6-13](#).



■ Fig. 6-13

A fringe tracking system for a seven-beam interferometer. In this system, only seven baselines are sampled, but they include all seven telescopes and will therefore provide all the necessary information required to phase lock all 21 possible baselines. Here the baselines sampled are beams 17, 12, 23, 34, 45, 56, and 67. The dispersion before the cameras provides the ability to perform group delay tracking


When the task of fringe tracking and imaging is separated, it is the fringe tracking system that determines the magnitude limit, and so the fringe tracker will most likely use an aperture-plane scheme. It must also have fast sample times and will most likely not use spatial filtering. It is not necessary in the case of the fringe tracker to be concerned about data calibration; it is only necessary to be able to detect fringe phase. As long as the fringe tracker locks the fringe phase, the imaging combiner can integrate on the fringes for all baselines. In this way, the sensitivity limitations of image-plane combination and spatial filtering with single-mode fibers can be overcome.

The fringe tracker can also take advantage of the layout of the interferometric array. Resolution scales with B/λ_0 and visibility amplitude will be large for smaller resolutions. Thus, since it is only necessary to sample a small number of baselines in the fringe tracker, it is best to ensure that these are the smallest baselines possible. This method is called *bootstrapping*. Moreover, the light is most often broken up by wavelength, so it is preferable to use a longer wavelength for fringe tracking in order to ensure the highest possible signal-to-noise ratio.

Alignment System – An optical interferometer is a very complex instrument containing large numbers of optomechanical systems that must be properly coaligned if it is to work at all. This requires careful alignment and frequent adjustments. By the very nature of long baseline interferometry, the components that require adjustment can be hundreds of meters away from the operator so a large number of remote actuators and sensors will be required.

Probably the most essential part of the alignment system is a method of injecting both laser and white light into the back end of the beam combiner, which then propagates through the entire optical system all the way out to the telescopes. Various pop-up targets, screens, and video cameras can then be used to ensure the laser hits all of the designated targets throughout the optical path. One way of creating these fiducials is to drill small holes in the center of some of the mirrors. The laser will have a shadow created by each of these holes and these must all overlap one another. It is also possible to place LEDs in these holes.

Once the laser, or white light, reaches the telescope, it can be retroreflected by a corner cube and sent all the way back into the laboratory and through the beam combiner. In telescopes with a secondary shadow, this corner cube can be placed in the middle of the secondary. This makes it possible to form fringes using the white light source and test all the internal detector systems and servos. It is also an excellent way of measuring the differential internal path lengths of each baseline as well as the path length changes introduced by each of the POPs.

Putting It All Together – An example full back end for a six-beam interferometer is shown in  Fig. 6-14. This layout includes tip/tilt detection, a single-mode fiber-based imaging beam combiner, a bulk optic aperture-plane fringe tracking combiner, and the necessary parts for injecting alignment laser and white light sources.

This is by no means the only, or even necessarily the best, way of designing an interferometer. Indeed, there is no “best” way. In the end, design decisions will be based on the intended science goals, on the personal preferences of the design team, and also to a large extent on the available budget. Nevertheless while each interferometer is unique, there are common attributes to them all, and there is little doubt that their design will continue to evolve with time.

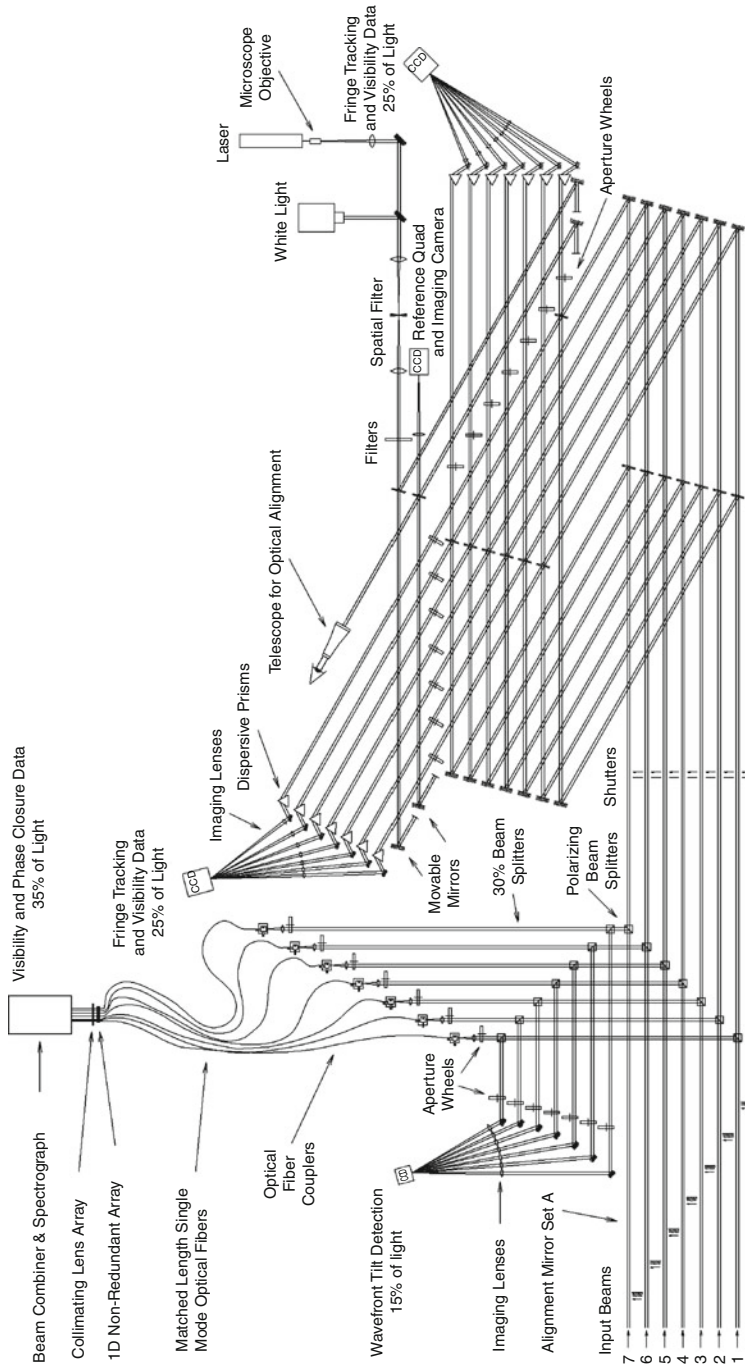


Fig. 6-14
A possible layout for a complete back end beam-combining system for a seven-beam optical interferometer

References

- Armstrong, J. T., Mozurkewich, D., Rickard, L. J., Hutter, D. J., Benson, J. A., Bowers, P. F., Elias, N. M., II, Hummel, C. A., Johnston, K. J., Buscher, D. F., Clark, J. H., III, Ha, L., Ling, L.-C., White, N. M., & Simon, R. S. 1998, The navy prototype interferometer. *ApJ*, 496, 550
- Baldwin, J. E., Boysen, R. C., Haniff, C. A., Lawson, P. R., Mackay, C. D., Rogers, J., St-Jacques, D., Warner, P. J., Wilson, D. M., & Young, J. S. 1998, Current status of COAST. *Proc SPIE*, 3350, 736
- Benisty, M., Berger, J.-P., Jocou, L., Labeye, P., Malbet, F., & Perraut, K. 2009, An integrated optics beam combiner for the second generation VLTI instruments. *A&A*, 498, 601
- Born, M., & Wolf, E. 2002, *Principles of Optics* (7th ed.; Cambridge/New York: Cambridge University Press)
- Breckinridge, J. B., McAlister, H. A., & Robinson, W. G. 1979, Kitt peak speckle camera. *Appl Opt*, 18, 1034
- Buscher, D. F. 1988, Optimizing a ground-based optical interferometer for sensitivity at low light levels. *MNRAS*, 235, 1203
- Bracewell, R. N. 2000, *The Fourier Transform and its Applications* (3rd ed.; Singapore: McGraw Hill Higher Education)
- Che, X., Monnier, J. D., & Webster, S. 2010, High precision interferometer: MIRC with photometric channels. *Proc SPIE*, 7734, 91
- Colavita, M. M. 1999, Fringe visibility estimators for the palomar testbed interferometer. *PASP*, 111, 111
- Colavita, M., Shao, M., & Staelin, D. H. 1987, Atmospheric phase measurements with the Mark III stellar interferometer. *Appl Opt*, 26, 4106
- Colavita, M. M., Hines, B. E., Shao, M., Klose, G. J., & Gibson, B. V. 1991, Prototype high speed optical delay line for stellar interferometry. *Proc SPIE*, 1542, 205
- Colavita, M. M., Wallace, J. K., Hines, B. E., Gursel, Y., Malbet, F., Palmar, D. L., Pan, X. P., Shao, M., Yu, J. W., Boden, A. F., Dumont, P. J., Gubler, J., Koresko, C. D., Kulkarni, S. R., Lane, B. F., Mobley, D. W., & van Belle, G. T. 1999, The palomar testbed interferometer. *ApJ*, 510, 505
- Colavita, M., Akeson, R., Wizinowich, P., Shao, M., Acton, S., Beletic, J., Bell, J., Berlin, J., Boden, A., Booth, A., & 53 other coauthors. 2003, Observations of DG Tauri with the Keck Interferometer. *ApJ*, 592, 83
- Coudé du Foresto, V., Ridgway, S., & Mariotti, J. M. 1997, Deriving object visibilities from interferograms obtained with a fiber stellar interferometer. *A&AS*, 121, 2
- Coudé du Foresto, V., Perrin, G., Ruilier, C., Mennesson, B. P., Traub, W., & Lacasse, M. G. 1998, FLUOR fibered instrument at the IOTA interferometer. *Proc SPIE*, 3350, 856
- Coulman, C. E. 1985, Fundamental and applied aspects of astronomical seeing. *Ann Rev Astron Astrophys*, 23, 19
- Creech-Eakman, M. J., Romero, V., Payne, I., Haniff, C., Buscher, D., Aitken, C., Anderson, C., Bakker, E., Coleman, T., Dahl, C., & 26 other coauthors. 2010, Magdalena Ridge Observatory Interferometer advancing to first light and new science. *Proc SPIE*, 7734, 11 (page numbers odd on ads)
- Davis, J., Lawson, P. R., Booth, A. J., Tango, W. J., & Thorvaldson, E. D. 1995, Atmospheric path variations for baselines up to 80 m measured with the Sydney University Stellar Interferometer. *MNRAS*, 273, 53
- Davis, J., Tango, W. J., Booth, A. J., ten Brummelaar, T. A., Minard, R. A., & Owens, S. 1998, The Sydney University Stellar Interferometer – I. The instrument. *MNRAS*, 303, 773
- Davis, S., Abrams, M. C., & Brault, J. W. 2001, *Fourier Transform Spectroscopy* (San Diego: Academic Press)
- Delplancke, F. 2008, The PRIMA facility phase-referenced imaging and micro-arcsecond astrometry. *New Astron Rev*, 52, 199
- Dyck, H. M., Benson, J. A., Carleton, N. P., Coldwell, C., Lacasse, M. G., Nisenson, P., Panasyuk, A., Papaliolios, C., Pearlman, M. R., Reasonberg, R. D., Truab, W. A., Xu, A., Predmore, C. R., Schloerb, F. P., & Gibson, D. M. 1995, First 2.2 micrometer results from the IOTA interferometer. *AJ*, 109, 378
- Elias, N. M. 2001, Optical interferometric polarimetry I. Foundation. *ApJ*, 549, 647
- Elias, N. M. 2004, Optical interferometric polarimetry II. Theory. *ApJ*, 611, 1175
- Ferrari, M., Lemaître, G. R., Mazzanti, S. P., Derie, F., Huxley, A., Lemerer, J., Lanzoni, P., Dargent, P., & Wallander, A. 2003, Variable curvature mirrors: implementation in the VLTI delay-lines for field compensation. *Proc SPIE*, 4838, 1155
- Fisher, M., Boysen, R. C., Buscher, D. F., Haniff, C. A., Seneta, E. B., Sun, X., Wilson, D. M. A., & Young, J. S. 2010, Design of the MROI delay line optical path compensator. *Proc SPIE*, 7734, 135

- Frederic, D. 2000, VLT delay lines: design, development and performance requirements. *Proc SPIE*, 4006, 25
- Goodman, J. W. 2005, *Introduction to Fourier Optics* (Greenwood Village, USA: Roberts and Company Publishers)
- Haguenauer, P., Alonso, J., Bourget, P., Brillant, S., Gitton, P., Guisard, S., Poupau, S., Schuhler, N., Abuter, R., Andolfato, L., & 32 other coauthors. 2010, The very large telescope interferometer: 2010 edition. *Proc SPIE*, 7734, 4 (page numbers odd on ads)
- Hale, D. D., Bester, M., Danchi W. C., Fitelson, W., Hoss, S., Lipman, E. A., Monnier, J. D., Tuthill, P. G., & Townes, C. H. 2000, The Berkeley Infrared Spatial Interferometer: a heterodyne stellar interferometer for the mid-infrared. *ApJ*, 537, 998
- Hinz, P. M., Connors, T., McMahon, T., Cheng, A., Peng, C. Y., Hoffmann, W., McCarthy, D., Jr., & Angel, R. 2004, Large Binocular Telescope Interferometer: the universal beam combiner. *Proc SPIE*, 5491, 787
- Horton, A. J., Buscher, D. F., & Haniff, C. A. 2001, Diffraction losses in ground-based optical interferometers. *MNRAS*, 327, 217
- Hutter, D. J., & Elias, N. M. 2003, Array metrology system for an optical long-baseline interferometer. *Proc SPIE*, 4838, 1234
- Ireland, M. J., Mérand, A., ten Brummelaar, T. A., Tuthill, P. G., Schaefer, G. H., Turner, N. H., Sturmman, J., Sturmman, L., & McAlister, H. A. 2008, Sensitive visible interferometry with PAVO. *Proc SPIE*, 7013, 63
- Kolmogorov, A. 1941a, On degeneration of isotropic turbulence in an incompressible viscous fluid. *C R Acad Sci URSS*, 31, 538, Reprinted 1961 in "Turbulence," Interscience Publishers, Inc., New York, Ed: Friedlander, S.K. and Topper, L
- Kolmogorov, A. 1941b, Dissipation of energy in the locally isotropic turbulence. *Compt Rend Acad Sci USSR*, 32, 16–18, Reprinted 1961 in "Turbulence," Interscience Publishers, Inc., New York, Ed: Friedlander, S.K. and Topper, L
- Labeyrie, A. 1974, Speckle interferometry and possible extensions. *Astron Astrophys Suppl*, 15, 463
- Labeyrie, A., Lipson, S. G., & Nisenson, P. 2006, *An Introduction to Optical Stellar Interferometry* (Cambridge: Cambridge University Press)
- Lawson, P. R. (ed.), 1997, *Selected Papers on Long Baseline Stellar Interferometry*, Proc. (Bellingham, WA: SPIE Press)
- Lawson, P. R. (ed.), 1999, *Principles of Long Baseline Stellar Interferometry: Course notes from the 1999 Michelson Interferometry Summer School* (Pasadena, CA: Jet Propulsion Laboratory), available from the Jet Propulsion Laboratory
- Lohmann, A. W., Weigelt, G., & Wirtitzer, B. 1983, Speckle masking in astronomy – tripple correlation theory and applications. *Appl Opt*, 22, 4028
- Michelson, A. A., & Pease, F. G. 1921, Measurement of the diameter of Alpha Orionis with an interferometer. *ApJ*, 53, 249
- Monnier, J. 2003, Optical interferometry in astronomy. *Rep Prog Phys*, 66, 789
- Monnier, J., & Allen, R. J., 2011, The article in the other book in this series describing how to reduce data from interferometers (Springer)
- Nightingale, N. S., & Buscher, D. F. 1991, Interferometric seeing measurements at La Palmer. *MNRAS*, 251, 155
- Noll, R. J. 1976, Zernike polynomials and atmospheric turbulence. *JOSA*, 66, 207
- Quirrenbach, A. 2001, Optical interferometry. *Ann Rev Astron Astrophys*, 39, 353
- Richardson, L. F. 1922, *Weather Prediction by Numerical Process* (Cambridge: Cambridge University Press), Reprinted 1965, Dover Pub. Inc., New York, Intro: Chapman, S.
- Roddier, F. 1981, The effects of atmospheric turbulence in optical astronomy. *Prog Opt*, XIX, 281
- Sheppard, C. R. J., & Hrynevych, M. 1992, Diffraction by a circular aperture: a generalization of Fresnel diffraction theory. *JOSA-A*, 9, 274
- Stomski, P. J., Jr., Le Mignant, D., Wizinowich, P. L., Campbell, R. D., & Goodrich, R. 2003, Compensation for differential atmospheric refraction in the W.M. Keck Observatory adaptive Optics system. *Proc SPIE*, 4839, 943
- Sturmman, J., ten Brummelaar, T. A., Sturmman, L., & McAlister, H. A. 2010, Dual three-way beam combiner at the CHARA array. *Proc SPIE*, 7734, 104
- Tango, W. J. 1990, Dispersion in stellar interferometry. *Appl Opt*, 29, 516
- Tango, W. J., & Twiss, R. Q. 1980, Michelson stellar interferometry. *Prog Opt*, XVII, 241
- Taylor, G. I. 1921, Diffusion by continuous movements. *Proc Lond Math Soc*, 2, 196–211, Reprinted 1961 in "Turbulence," Interscience Publishers, Inc., New York, Ed: Friedlander, S.K. and Topper, L.
- Taylor, G. I. 1935, *Statistical Theory of Turbulence* (London: Royal Society), Reprinted 1961 in "Turbulence," Interscience Publishers, Inc., New York, Ed: Friedlander, S.K. and Topper, L.
- ten Brummelaar, T. A. 1994, The CHARA beam combiner design. *Proc SPIE*, 2200, 140

- ten Brummelaar, T. A. 1995, Differential path considerations in optical stellar interferometry. *Appl Opt*, 34, 2214
- ten Brummelaar, T. A. 1996, Correlation measurement and group delay tracking in optical stellar interferometry with a noisy detector. *MNRAS*, 285, 135
- ten Brummelaar, T. A., & Tango, W. J. 1994, A wavefront tilt correction servo for the Sydney University Stellar Interferometer. *Exp Astron*, 4, 297
- ten Brummelaar, T. A., Bagnuolo, W. G., Jr., & Ridgway, S. T. 1995, Strehl ratio and visibility in long baseline interferometry. *Opt Lett*, 20, 521
- ten Brummelaar, T. A., McAlister, H. A., Ridgway, S. T., Bagnuolo, W. G., Jr., Turner, N. H., Sturmman, L., Sturmman, J., Berger, D. H., Ogden, C. E., Cadman, R., Hartkopf, W. I., Hopper, C. H., & Shure, M. A. 2005, First results from the CHARA array II: description of the instrument. *ApJ*, 628, 453
- Traub, W. A. 1988, *ESO Conf Workshop Proc*, 29, Part 2, 1029
- Tyson, R. K. 2000, *Adaptive Optics Engineering Handbook* (New York: Marcel Dekker)
- Van Cittert, P. H. 1934, Die Wahrscheinliche Schwingungsverteilung in Einer von Einer Lichtquelle Direkt Oder Mittels Einer Linse Beleuchteten Ebene. *Physica*, 1, 201
- Vergnole, S., Kotani, T., Perrin, G., Delage, L., & Reynaud, F. 2005, Calibration of silica fibers for the Optical Hawaiian Array for Nanoradian Astronomy (OHANA): temperature dependence of differential chromatic dispersion. *Opt Commun*, 251, 115
- Willez, J., Akeson, R., Colavita, M., Eisner, J., Ghez, A., Graham, J., Hillenbrand, L., Millan-Gabet, R., Monnier, J., Pott, J.-U., & 17 other co-authors. 2010, ASTRA: astrometry and phase referencing astronomy on the Keck interferometer. *Proc SPIE*, 7734, 30
- Zernike, F. 1938, The concept of degree of coherence and its application to optical problems. *Physica*, 5, 785

7 Submillimeter Telescopes

Thomas G. Phillips¹ · Stephen Padin² · Jonas Zmuidzinas³

¹Physics, Mathematics & Astronomy, California Institute of Technology, Pasadena, CA, USA

²California Institute of Technology, Pasadena, CA, USA


³George W. Downs Laboratory of Physics, California Institute of Technology, Pasadena, CA, USA




1	Introduction	284
2	Submillimeter Detection	289
2.1	Heterodyne Detectors	290
2.1.1	Hot Electron Bolometers	290
2.1.2	SIS Detectors	292
3	Telescopes	296
3.1	Optics	296
3.1.1	Cassegrain and Gregory Telescopes	296
3.1.2	Plate Scale	300
3.1.3	Other Telescope Designs	301
3.1.4	Chopping and Scanning	301
3.1.5	Scattering and Loss	302
3.1.6	Optical Components	302
3.2	Structure and Mechanics	302
3.2.1	Pointing and Wavefront Errors	303
3.2.2	Primary Mirror Support	304
3.2.3	Mirror Control	306
3.2.4	Telescope Mount	307
3.3	Alignment	309
	References	311

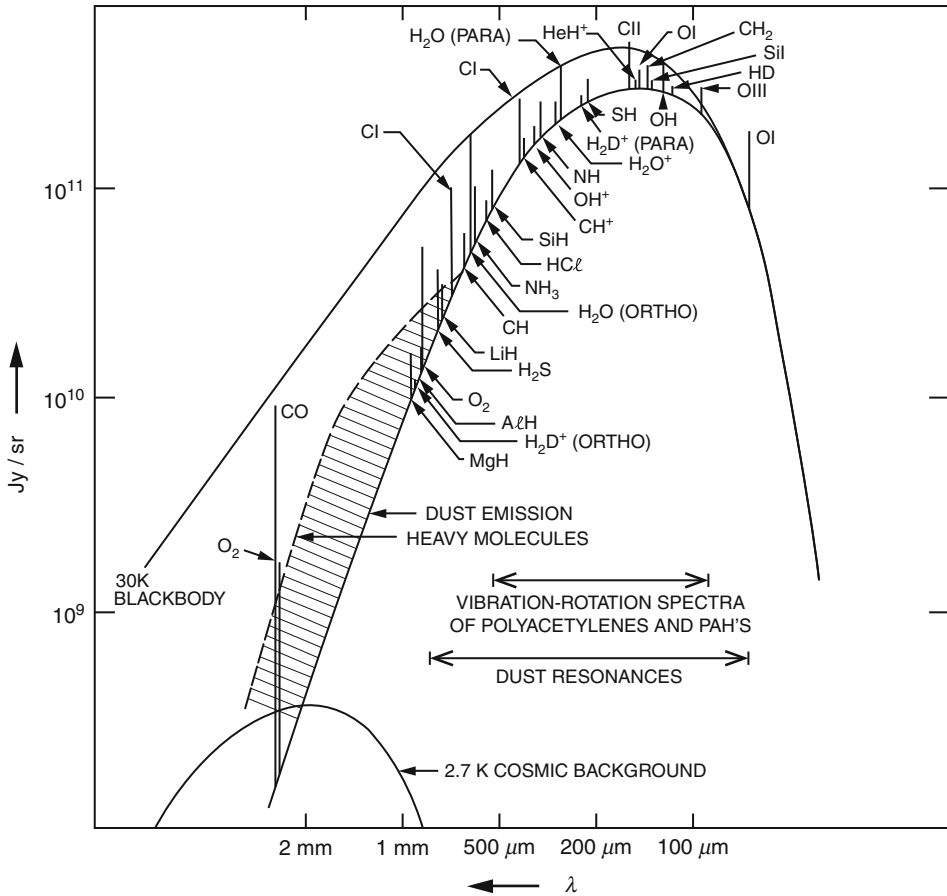
List of Abbreviations: *CFRP*, Carbon fiber reinforced plastic; *CMB*, Cosmic microwave background; *FOV*, Field of view; *PSF*, Point spread function; *RC*, Ritchey-Chrétien

1 Introduction

The submillimeter band is a critical one for astronomy. It contains spectral and spatial information on very distant newly formed galaxies and on the early stages of star formation within gas clouds. Yet it is one of the few regions of the electromagnetic spectrum still to be made fully available to astronomy. This is in part due to the general difficulties of construction of detectors, receivers, and telescopes for these wavelengths and in part to the attenuating nature of the Earth's atmosphere. In recent years, optical style telescopes have become available, either on high mountain sites, or in the case of the NASA Kuiper Airborne Observatory (KAO) or Stratospheric Observatory for Infrared Astronomy (SOFIA) on board a high-altitude airplane. The James Clerk Maxwell telescope at 15 m and the Caltech Submillimeter Observatory (CSO) telescope at 10.4 m are both large enough to have developed the field. However, the ESA satellite Herschel has now provided the required space platform for complete spectral coverage and the Atacama Large Millimeter/Submillimeter Array (ALMA) the high spatial resolution, aperture synthesis, high-sensitivity platform.

On the whole, the emission strength is low in the submillimeter for astronomical objects. The electronic processes which provide strong emission in the radio fade away at high frequencies, and the thermal emission from cold objects is relatively weak, particularly when highly redshifted. An overall view of the spectrum of a dense interstellar cloud in our own galaxy provides a sense of the spectroscopic requirements.  *Figure 7-1* shows a schematic representation of the likely emission from a typical star-forming cloud in the galaxy. The cloud dust and gas temperatures are assumed to be 30 K. A black body curve provides an upper bound to the emission strength, apart from possible maser action which is ignored here. Dust is optically thick at short wavelengths, and the continuum emission from it lies on the black body curve, but drops well below at long wavelengths where the dust is optically thin. At millimeter wavelengths the gas spectrum is dominated by emission due to the rotation spectrum of heavy molecules. The strongest of these by far is that of the carbon monoxide (CO) molecule which is so abundant that it can have optical depths of ~ 100 , even though it has a dipole moment of only ~ 0.12 debye. Low-lying transitions of CO usually reflect the gas temperature and are represented in the figure as reaching the black body curve, at least for the line centers. However, for high values of the rotational quantum number (J), the molecules tend to be deexcited and emit less strongly, so the line intensities drop below the black body curve.

An actual example of part of the millimeter band spectrum of the Orion Cloud (OMC-1) with the continuum removed is shown in  *Fig. 7-2*. This was taken with one of the Leighton Caltech telescopes using an SIS (see  *Sect. 2*) receiver (Blake et al. 1987). The result was the detection of about 1,000 lines which were identified as various transitions of about 30 molecules, varying in complexity from CO to dimethyl ether (CH_3OCH_3). The spectrum observed was about 60 GHz in width which permitted multiline observations of the molecules so necessary for secure identification. It is clear that providing a large instantaneous bandwidth is a primary requisite of a millimeter or submillimeter receiver system. Much higher spectral resolution is available. An expanded region is shown in  *Fig. 7-3*. It turns out that several different spatial

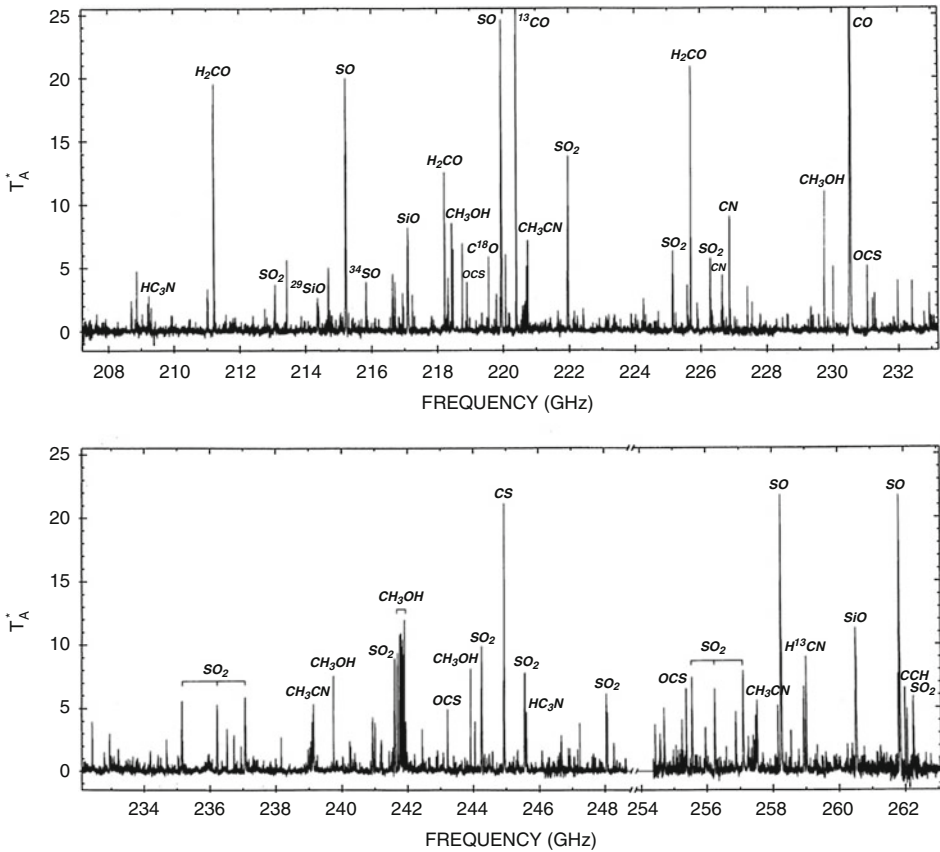


■ Fig. 7-1

The anticipated spectrum of a 30 K cloud, showing the dust emission, the heavy molecule line forest, and, at high frequencies, the fast rotating light hydrides (Phillips and Keene 1992)

objects, within the one telescope beam, can be discriminated by their distinctive line shapes. This emphasizes a second requirement for interstellar spectroscopy, namely, very good spectral resolution.

Returning to [Fig. 7-1](#), as we move from the millimeter band, containing many lines of heavy molecules, to the submillimeter, we find that the spectrum is dominated by light molecules containing hydrogen. These light molecules have small moments of inertia and therefore rotate fast, so their fundamental rotation lines appear in the submillimeter. Heavy molecule lines become less important due to the de-excitation suffered by a molecule in a high J energy level, well above the ground state. The postulation of [Fig. 7-1](#) has been verified by the space project, HIFI, on the Herschel telescope orbiting at the L2 point. Many transition lines were recently discovered by the HIFI instrument, under the water lines which prevent observations from even high mountain sites. An example is HF ([Fig. 7-4](#)) and water itself was seen, initially for weak lines for which the atmosphere is not opaque, mostly from the KAO. More recently,



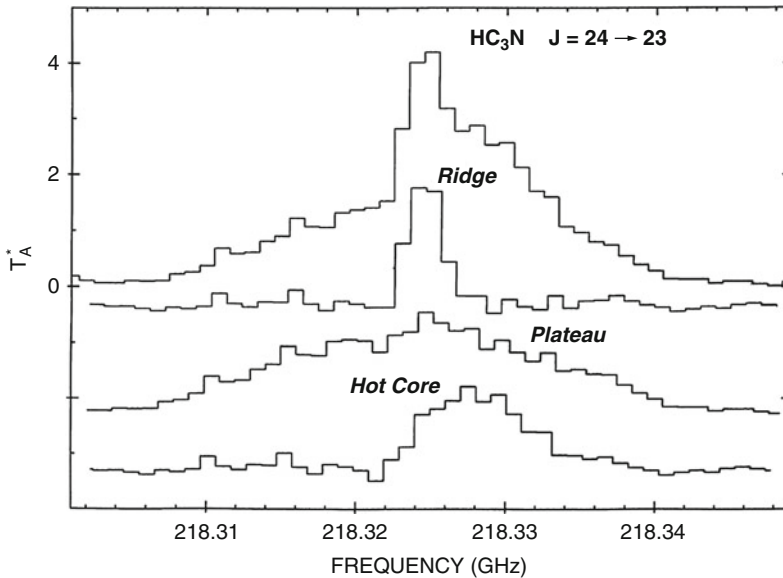
■ Fig. 7-2

Part of the heavy molecule spectrum as seen from the first Leighton telescope in Blake's thesis

it was seen from space probes such as SWAS and finally, definitively, with HIFI. Although it is clear from [Fig. 7-1](#) that there is a tremendous quantity of astrophysical information contained in the submillimeter band, it has proved hard to obtain, even for our own galaxy, because of the interfering effects of the Earth's atmosphere.

[Figure 7-5](#) shows the transmission for good weather (about 1 mm of precipitable water) at the 4,200-m level on Mauna Kea for the submillimeter band. There are three useful regions: the radio/millimeter region from 0 to 300 GHz which is almost completely transparent, apart from a few O_2 and H_2O lines, and the two wide submillimeter windows at roughly 650 GHz ($\sim 450 \mu m$) and 850 GHz ($\sim 350 \mu m$). The rest of the spectrum is blacked out by water lines, apart from some narrow windows near 400 GHz. The atmosphere is completely black at shorter wavelengths until the mid-infrared windows at $\sim 30 \mu m$ are reached.

At airborne altitudes, say 12,000 m for SOFIA, the transmission is much better, but also not perfect. Several of the lines of [Fig. 7-1](#) were initially detected with early versions of submillimeter detectors from the KAO, such as the fundamental rotation lines of NH_3 (Keene et al. 1983) and HCl (Blake et al. 1985). Many of the high-frequency lines of CO (Phillips et al. 1980a) were observed and some lines of OH and CH (Stacey et al. 1987; Storey et al. 1981).



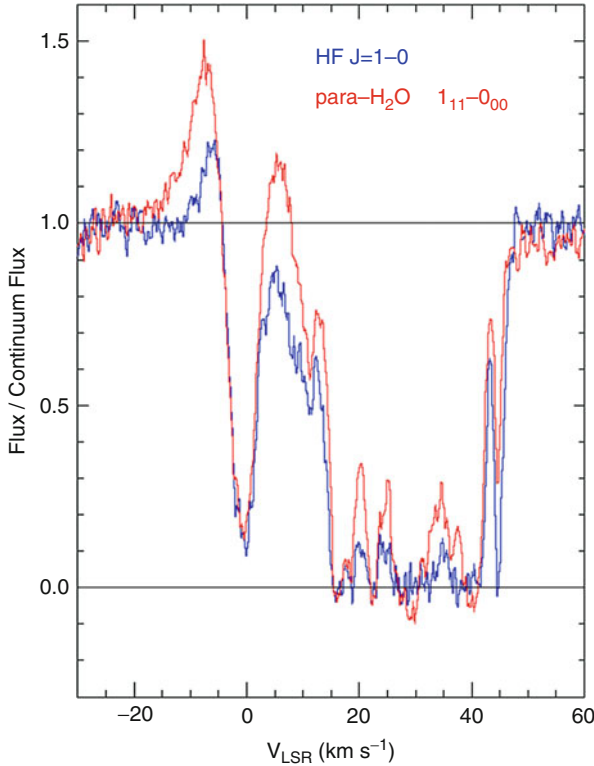
■ Fig. 7-3
The Orion “hot core,” “Plateau” (outflow), and quiescent “ridge”

Also the atomic fine structure lines of carbon and ionized carbon have been detected by various techniques (Jaffe et al. 1985; Phillips and Huggins 1981; Russell et al. 1980).

With the advent of the mountain-sited submillimeter telescopes and new space-based telescopes, the quality of submillimeter spectroscopy has surpassed that of the millimeter spectrum of Fig. 7-2. An example of the modern line surveys (HIFI) is shown in Fig. 7-6. The data is taken double sideband but deconvolved to single sideband. In actuality, these spectra include common molecules with so many lines that they have to be subtracted in order to see the underlying spectrum. They are designated as “weeds” (an example is CH_3OH).

Being situated between the radio and infrared fields, submillimeter astronomy has naturally borrowed its techniques from these more established areas. The radio is generally considered to be a thermally dominated regime, i.e., $kT > h\nu$; the thermal energy is greater than the photon energy. By contrast, the infrared or optical regimes are in the limit where quantum energies dominate thermal. In the submillimeter, neither thermal effects nor quantum effects can necessarily be assumed negligible and so both must be considered when designing detectors. The two basic methods of detection to be compared are the infrared technique of direct detection, such as bolometric or photodetection, and the radio technique of heterodyne mixing which has been so successful in the millimeter and submillimeter bands. We can make the comparison for a range of required spectral resolutions and astronomical conditions (Phillips 1988). The radio approximation $kT \gg h\nu$ is often used, even though it is not fully valid.

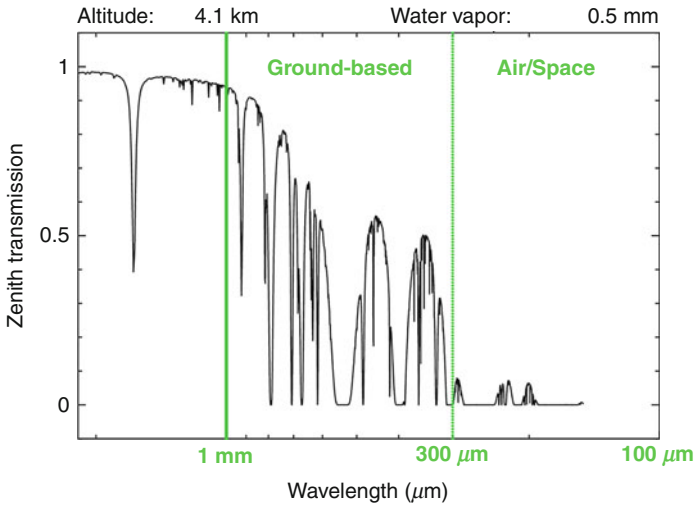
There are clearly two types of science that can be carried out in the submillimeter; continuum observation of dust and high spectral resolution measurements of atomic and molecular line features. The importance of the dust observation became clear when Michael Rowan-Robinson (Rowan-Robinson et al. 1991) found that the IRAS object F10214 was at a large distance ($z = 2.3$). The bolometers in use were the doped Ge type, developed by Frank Low (1961).



■ Fig. 7-4

HF is a strongly bonded, fast-rotating molecule, which is often seen in absorption, and in hot dense regions in emission (Neufeld et al. 2010) and a water absorption spectrum for comparison. The spectra are very similar as seen when the frequencies of observation are shifted to allow this comparison

➤ Figure 7-7 shows a photograph of Phillips and Huggins with the initial InSb hot electron bolometer (see ➤ Sect. 2) receiver mounted at the prime focus of the 200-in. telescope (Phillips et al. 1977). Due to the water and oxygen in the Earth's atmosphere, work from the ground becomes very difficult in the submillimeter, and, as a result, some initial effort was switched to operation on the KAO. Flying at an altitude of 40,000 feet effectively reduced the attenuation by the atmosphere to a reasonable value and allowed observations at much higher frequencies. These included the detection of the CO (4–3) line at 460 GHz (Phillips et al. 1980a); the ground state fine-structure line of atomic carbon (C I) at 492 GHz (Phillips et al. 1980b); the fundamental rotational line of ammonia, NH₃ (1–0) at 572 GHz (Keene et al. 1983); the fundamental rotational line of HCl (1–0) at 626 GHz (Blake et al. 1985); etc. The telescope was small (91.5 cm) but served the purpose of allowing the discovery of many of the molecular and atomic species present in the dense interstellar medium. This work has been now superseded by the SIS receivers (see ➤ Sect. 2) on HIFI/Herschel and SOFIA. SIS receivers also were chosen for ALMA, probably the outstanding astronomy instrument of the decade.



■ Fig. 7-5

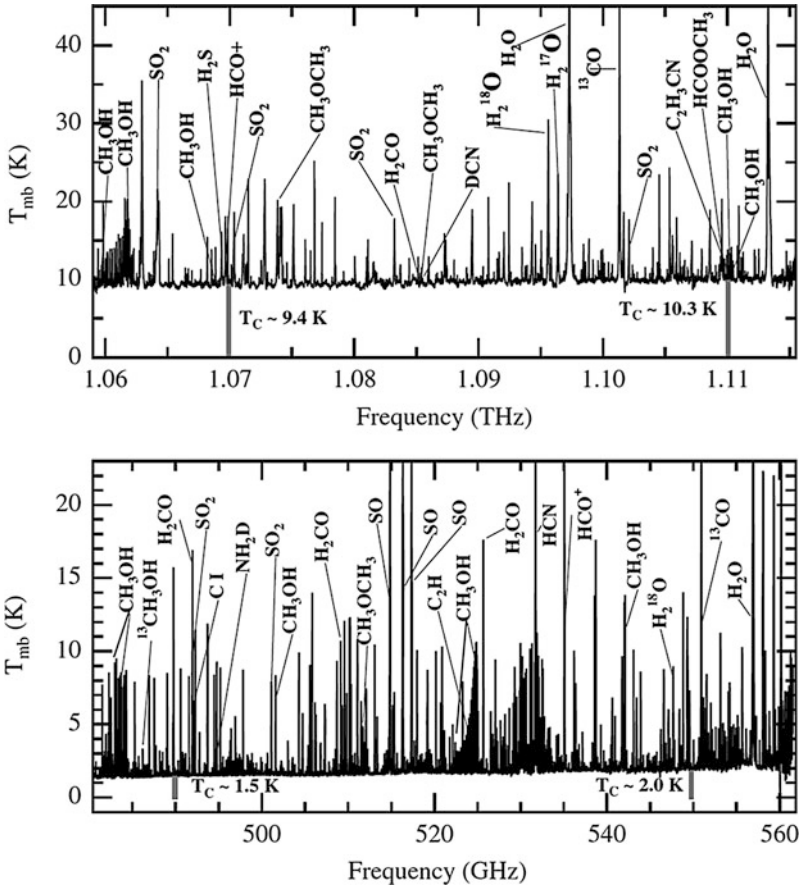
The Earth's atmospheric transmission in the submillimeter band as seen from the CSO. A suitable platform for the various frequency ranges is shown in *green*

2 Submillimeter Detection

There are two basic techniques for millimeter and submillimeter detection. The first is direct detection with a bolometer or photodetector device, and the second is heterodyne detection using a mixer device. Although modern bolometers are more sensitive than mixers, it is still necessary to use mixers where it is required to keep track of phase, e.g., for interferometers, or where very high spectral resolution is needed. Bolometer direct detection requires a detection element such as resistance or reactance which is a function of temperature. If the element has a sufficiently rapid function of temperature that the signal voltage across the device is very much larger than all locally generated noise voltages (e.g., Johnson noise or thermal fluctuation noise), then the signal to noise ratio (S/N) is determined by the background noise. By contrast, the heterodyne process is limited by quantum noise ($\sim h\nu/k$) which is a linear function of frequency. Thus, in general, one can say that direct detection is preferred at infrared wavelengths, and heterodyne techniques are preferred in the radio. In a given case, where noise is determined by the detector element, a comparison will depend on the channel width ($\Delta\nu$), the noise equivalent power (NEP) of the direct detector, and the noise temperature of the heterodyne instrument (T_N) as:

$$\frac{(S/N)_{\text{heterodyne}}}{(S/N)_{\text{direct}}} = \frac{\text{NEP}}{2kT_N\sqrt{\Delta\nu}}$$

In the case where the background is dominant, this ratio depends only on the root of the ratio of channel widths (Phillips 1988).



■ Fig. 7-6

A line survey of Orion using HIFI. The spectrum is amazingly complicated

2.1 Heterodyne Detectors

2.1.1 Hot Electron Bolometers

Any power detector, such as a bolometer, acts to detect the square of the total electric field and can therefore act as a mixer. It comes in two forms, semiconducting and superconducting. First, came the semiconducting device (Phillips and Jefferts 1973). This surprising simple detector was a rod of indium antimonide (InSb) mounted in the E-field of a full-height waveguide with a directly coupled baseband IF amplifier. The device was matched to the waveguide using an E-H tuner in front and a backshort behind. The waveguide device was mounted in a liquid helium cryostat resulting in the first very low temperature mixer receiver for millimeter or submillimeter astronomy. The way in which this hot electron bolometer worked was to absorb the incoming photons via the electron gas and make use of the temperature variation of the electron resistance due to the fact that the resistance was caused by impurity scattering of the



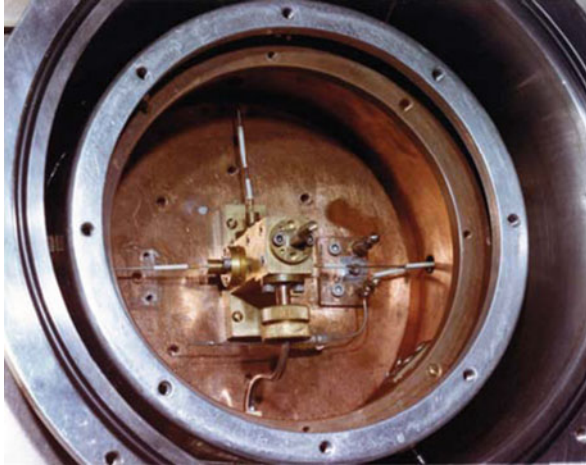
■ Fig. 7-7

Phillips and Huggins at the prime focus of the Palomar 200'' with the InSb hot electron bolometer

Rutherford type. The scattering is less effective as the electrons speed up resulting in a temperature dependence of the resistance. The original mounting scheme showing a scalar feedhorn, a 20-dB coupler for the local oscillator, and various waveguide tuners is displayed in [▶ Fig. 7-8](#).

The simple InSb detector really opened up the millimeter and submillimeter field of interstellar spectroscopy in spite of the very small IF bandpass (~ 1 MHz). This bandpass is set by the relaxation rate of the hot electrons excited by the DC bias. If the astronomical spectrum of the source was needed, the local oscillator had to be swept in frequency and the single pixel back end provided the spectrum. If a single channel map were desired, the local oscillator was fixed and the telescope swept over the desired area of the sky. [▶ Figure 7-9](#) shows a photograph of the InSb receiver mounted at the focus of the NASA Kuiper Observatory, used for frequencies from 400 to 600 GHz.

A superconducting version of the hot electron bolometer has been developed more recently (Gershenson et al. 1990). This is a device in which a small area (submicron) of a thin superconducting ribbon (a few nm) is heated by the DC bias and the local oscillator power (McGrath et al. 1997) and the IF passband is determined by the phonon interactions which cool the electrons (see [▶ Fig. 7-10](#)). An alternative mixer has been constructed using the electron diffusion out of the hot region to provide the cooling (Prober 1993) (see [▶ Fig. 7-11](#)). Both styles of HEB have fast electron cooling which allows an IF of about 2 GHz. The superconducting material is



■ Fig. 7-8

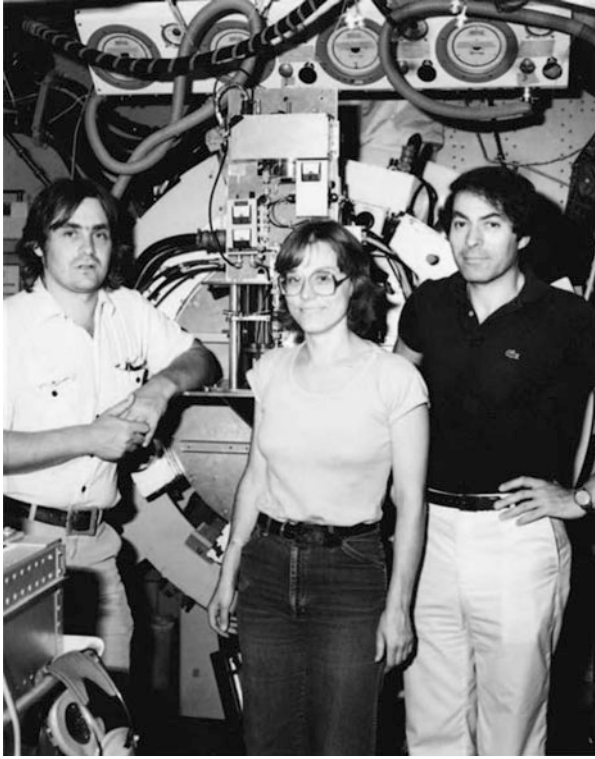
The waveguide mount for the InSb hot electron bolometer. The device is matched to the waveguide by an E-H tuner and a backshort

either Nb or NbN. The upper end of the frequency response of the devices is determined by the available local oscillator power. Probably the most effective application, to date, is in the HIFI instrument on Herschel where the local oscillator power is available to 1.9 THz. This is achieved by means of a synthesizer at ~ 30 GHz which is multiplied up by a series of multipliers (Pearson et al. 2000).

2.1.2 SIS Detectors

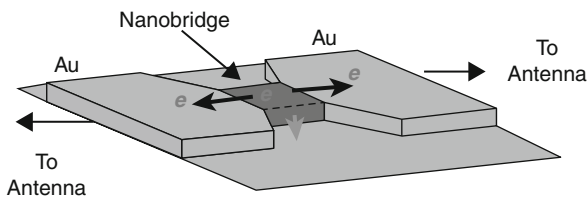
Superconductor-insulator-superconductor (SIS) receivers were developed independently at Bell Labs for 115 GHz mixing (Dolan et al. 1979) and for 36 GHz (Richards et al. 1979) at Berkeley. They have proved to be the most effective receivers in the submillimeter band of 100–800 GHz and have been used in nearly all receivers in telescopes which work in this range. Although considerable success had been achieved by the semiconductor hot electron bolometer receiver in opening up a new, complex, and interesting field, it was clear that a better receiver was needed as the 1-MHz baseband was unacceptable for a general purpose instrument. The alternative at that time was to attempt to operate the Schottky diode receivers at submillimeter wavelengths. Although this was possible, they were rather noisy, so an improved low-temperature receiver was required. The low-temperature solution to the receiver problem came about from the intense rivalry between AT&T and IBM in their competition to develop a superconducting computer. Neither company was successful, mainly due to the instability of the lead alloy superconducting tunnel junction electrodes. However, they did develop a high-frequency switching system using superconducting tunnel junctions which could be used as fast millimeter or submillimeter wave detectors. Bell Labs supported this and the necessary lithography.

The effect in use is the photon-assisted quasiparticle tunneling discovered by Dayem and Martin at Bell Labs (1962), but passed over as a device due to the more glamorous Josephson



■ Fig. 7-9

From the left: Chas Beichman, Jocelyn Keene, and Tom Phillips with the InSb receiver mounted at the focus of the 91.5-cm telescope of the Kuiper Airborne Observatory (KAO), operated by NASA



■ Fig. 7-10

Basic geometry of the superconducting HEB mixer device

junction discovered soon afterward which, however, eventually proved too noisy. The physics of this low-temperature, quasiparticle tunneling superconducting device is related to that of the photodetectors used in the optical and infrared where the bandgap of semiconducting materials of about 1 eV is suitable to allow absorption of optical/IR photons. Equivalently, the superconducting energy gap of typically 1 or 2 meV is suitable for photo-detection in the millimeter and submillimeter involving the breaking of Cooper pairs. The superconducting devices are now known by their structure, “superconductor-insulator-superconductor” (SIS), rather

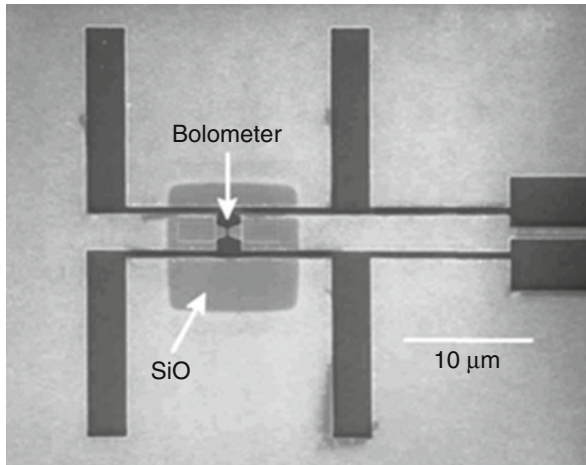


Fig. 7-11

SEM of Nb HEB mixer-embedding circuit showing twin-slot antenna and CPW lines. Inset shows area around the submicron device

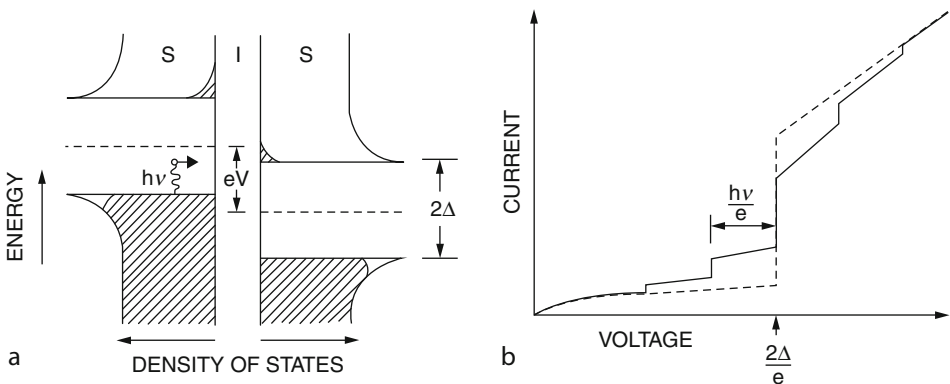


Fig. 7-12

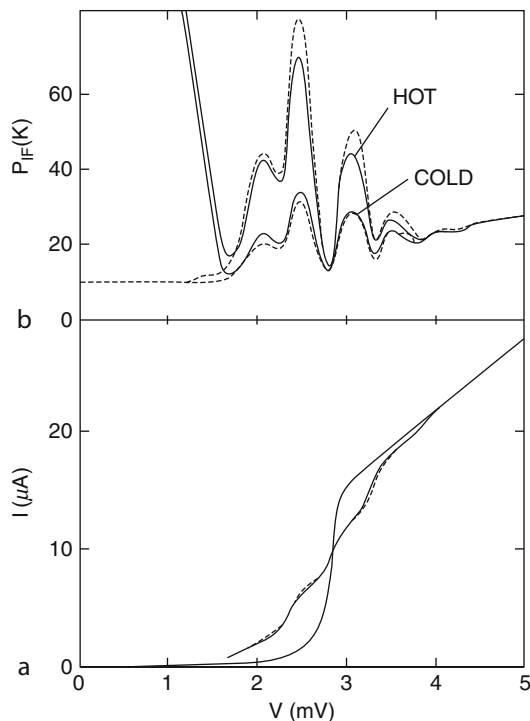
(a) Diagram of the superconducting bandgap and photon-assisted tunneling. (b) Idealized current as a function of the voltage. Dotted line is without any radiation and the solid line is with local oscillator radiation applied

than their physics since the SIS structure supports both quasiparticle and Josephson tunneling. This device, using the lead alloy materials, was mounted on a telescope for the first time in 1979 (Phillips and Woody 1982). The lead alloy materials were notoriously unstable both chemically and physically and were eventually replaced by niobium which is acceptably stable. During this time, an excellent theory of the operation of the SIS detector was developed by John Tucker (1979).

Early work on quasi-particle tunneling at Bell Labs demonstrated photon assisted tunneling which is revealed by step structure in the current vs. voltage at a spacing of $h\nu/e$ when local oscillator (LO) radiation at a frequency ν is applied (Dayem and Martin 1962). [Figure 7-12](#)


shows a diagram of the superconducting bandgap and photon-assisted tunneling for an SIS junction along with an idealized current vs. voltage characteristic. The quasiparticle tunneling description predicted the current vs. voltage characteristics as a function of the LO drive level, and this in turn was used to develop a phenomenological prediction of the mixer conversion efficiency (Phillips and Woody 1982).

The steps in the current vs. voltage relationship were seen in the early mixers and are shown in [Fig. 7-13](#) for an SIS receiver operating at 115 GHz. The hot and cold response does not look impressive by today's standards but achieving even this performance for the first-generation devices demonstrated the excellent potential for SIS mixers at millimeter and submillimeter wavelengths. A critical issue for SIS mixers was verification that it was the quasiparticle tunneling and not Josephson-pair tunneling that was responsible for the IF output power. A magnetic field is very effective at suppressing the effects of Josephson-pair tunneling. As seen in [Fig. 7-13](#), the IF output power near 0 v was decreased dramatically when a magnetic field was applied, but the output power at the photon step at the 2.3-mV gap voltage was only slightly changed by the application of a magnetic field. The conversion efficiency and noise temperature



■ Fig. 7-13

(a) Current vs. voltage characteristic without and with 115 GHz LO applied. (b) IF output power with the receiver looking at a 300 K room temperature load and at a 77 K liquid nitrogen load with the LO on. The solid line is without any magnetic field applied and dashed line is with a magnetic field applied. The photon-assisted tunneling is clearly revealed by the modulation in IF power as a function of the bias voltage (Woody 2009)

actually improved with the application of the magnetic field.  [Figure 7-12](#) also shows that even at 115 GHz the voltage step size is a significant fraction of the superconducting bandgap voltage. As these devices were pushed to higher frequencies, the Josephson-pair tunneling became more troubling, and new alloy systems needed to be developed for THz receivers.

Tucker's development of the full quantum mechanical description of quasi-particle tunneling and mixing in SIS tunnel junctions was a major advance that enabled a complete description and prediction of both the conversion efficiency and noise (Tucker 1979). This theory showed that the performance of SIS heterodyne mixers is limited only by the added noise imposed by quantum mechanics (Caves 1982; Wengler and Woody 1987). The mixers have conversion gains greater than unity, exceeding the limits for classical heterodyne mixing. This theoretical work was quickly incorporated into the receiver design process resulting in well-engineered astronomical receivers. Receiver performance within a factor of a few of the quantum limit for heterodyne mixing became feasible.

The critical component in SIS receivers is the tunnel junction. The current density, area, gap voltage, leakage current, as well as the stability of the thin film materials used are all important for producing reliable astronomical SIS receivers with good performance.


The widespread adoption of the SIS receivers on radio telescopes required the development of more reliable thin film systems. Most SIS receivers in the millimeter and low-frequency submillimeter band now utilize devices with Nb electrodes and Al-oxide tunnel barriers. This system produced high-quality devices with predictable and stable design parameters with the tuning structures and matching circuits fabricated on the chips along with the SIS device. Excellent performance is achieved throughout the millimeter and submillimeter bands (Kooi et al. 1992).

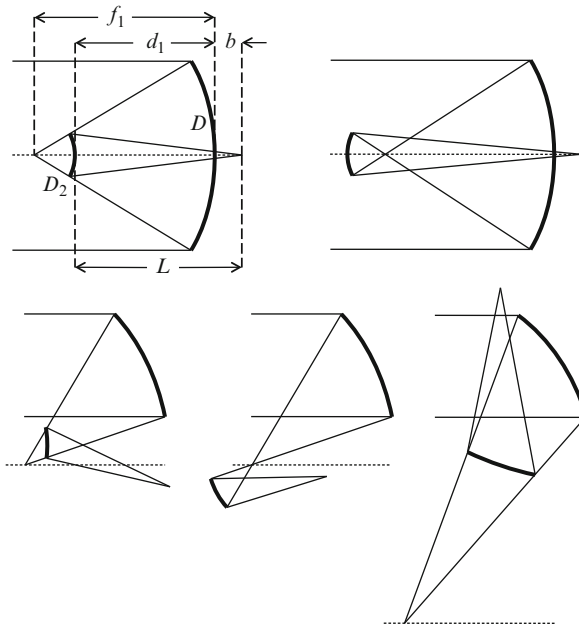
3 Telescopes

3.1 Optics

Most submillimeter telescopes are two-mirror reflecting telescopes (Baars 2007; Korsch 1991) similar to those used at optical and IR wavelengths (Wilson 1996, 1999). Reflecting surfaces have low loss, and the folded optical path results in a compact design, which is an important advantage for the telescope structure. The following sections give an overview of the most common designs. The emphasis here is on ground-based telescopes, but some examples of space telescopes are included.

3.1.1 Cassegrain and Gregory Telescopes

Submillimeter telescopes are usually on-axis Cassegrain or Gregory designs. Optical layouts for these telescopes are shown in  [Fig. 7-14](#). The classical Cassegrain and Gregory forms both have a paraboloidal primary (conic constant $b_1 = -1$). In the Cassegrain form, a convex hyperboloidal secondary (conic constant $b_2 < -1$) relays the inaccessible image at prime focus to the Cassegrain focus, which is usually behind the primary. A flat tertiary is often included to provide a Nasmyth focus on the elevation axis. The Cassegrain is the shortest of the two-mirror designs, so it is widely used. Examples of submillimeter classical Cassegrain telescopes include



■ Fig. 7-14

Cassegrain (top left) and Gregory (top right) telescopes, their off-axis variants (bottom left and center), and the crossed Mizuguchi-Dragone (bottom right). The horizontal dotted line in each diagram is the optical axis of the primary. D and D_2 are the diameters of the primary and secondary mirrors

CSO, JCMT, SMA, SMT, IRAM, APEX, ALMA, and Herschel (see ► Table 7-1). In the Gregory form, the secondary is a concave ellipsoid (conic constant $-1 < b_2 < 0$), mounted just beyond prime focus. The Gregory is longer, but its concave secondary is easier to test.

Submillimeter telescopes usually have a fast primary to keep the secondary support structure small. A primary focal ratio ≤ 0.4 minimizes the size and cost of the enclosure because the volume swept out by the telescope is set by the telescope diameter. Given the primary focal length, f_1 , and a back focal distance, b , chosen to give reasonable access to instruments, the focal length of the telescope, f , or secondary magnification, $m_2 = f/f_1$, or secondary diameter, D_2 , fixes the design (Wilson 1996). Small blockage generally requires $D_2 < D/10$, but if D_2 is too small, the focal length of the telescope becomes unreasonably large. The layout is specified by the secondary to image separation

$$L = |f| \frac{D_2}{D} = f \frac{1 - b/f_1}{1 - m_2}, \quad (7.1)$$

and the primary to secondary separation

$$d_1 = b - L = f_1 \frac{b/f_1 - m_2}{1 - m_2}, \quad (7.2)$$

and the secondary surface is specified by its radius of curvature

$$r_2 = \frac{2L}{1 + m_2}, \quad (7.3)$$

Table 7-1
Submillimeter telescopes^a

Telescope	Location	Altitude m	Diameter m	Surface error $\mu\text{m rms}$	Configuration ^b	Segments and support ^c	Enclosure
CSO (Leighton 1978)	Hawaii	4,200	10.4	13	Cassegrain	Al/Ah/Al, steel truss	Al dome, vertical shutter
JCMT (Baars 2007)	"	4,200	15	24	"	Al/Ah/Al, steel truss	Cylinder, sliding doors
SMA (Blundell 2007)	"	4,200	8 × 6	13	"	mAl, CFRP truss	No
SMT (HHT) (Baars 2007)	Arizona	3,300	10	15	"	CFRP/Ah/CFRP, CFRP truss	Box, hinged doors
IRAM (Baars et al. 1994; Guilloleau et al. 1992; Morris et al. 2009)	Spain	2,850	30	50	"	Al/Ah/Al, steel truss	No
	France	2,550	6 × 15	61	"	CFRP/Ah/CFRP, steel/CFRP truss	No
ALMA (Mangum et al. 2006; Saito 2011)	Chile	5,100	54 × 12	25	"	mAl, CFRP/Ah/CFRP box	No
APEX (Gusten et al. 2006)	"	5,100	12 × 7	20	"	eNi/Ah/eNi, CFRP plates mAl, steel truss	No
NANTEN2	"	4,900	4	17	"	"	No
KOSMA ^d (Kramer et al. 1998)	Switzerland	3,100	3	20	"	mAl, CFRP	Dome, horizontal shutter
AST/RO ^e (Stark, et al. 2001)	South Pole	2,800	1.7	30	OAG	mAl, CFRP truss	Dome, horizontal shutter
SPT (Carlstrom et al. 2011)	"	2,800	10	9	"	CFRP	Cloth dome, clamshell
ACT (Fowler et al. 2007; Swetz et al. 2011)	Chile	2,800	6	20	"	mAl, CFRP box	No
WMAP (Page et al. 2003)	L2	5,190	0.7	31	"	mAl, Al	Ground shield
Planck (Tauber et al. 2010)	L2		1.5	76	"	CFRP/KOREX/CFRP	
Herschel (Pilbratt et al. 2010)	L2		3.5	50	"	CFRP	
				2.5	Cassegrain	SIC	

^aThis list also includes some millimeter-wave telescopes to illustrate design options with low blockage

^bOAG = off-axis Gregory

^cAl/Ah/Al indicates a segment with Al facesheets and an Al honeycomb core. mAl/ machined Al, eNi electroformed Ni

^dMoved to Tibet in 2010

^eDecommissioned in 2005

Table 7-2

Sign conventions

Sign	Cassegrain	Gregory
+ve	f, b, L	b, L, m_2
-ve	f_1, d_1, m_2	f, f_1, d_1

Table 7-3

Third order aberrations for classical Cassegrain and Gregory telescopes (Schroeder 2000).

Aberration	Angular spot size ^a	FOV/ (λ/D)
Sagittal coma	$\theta/16F^2$	$32F^2$
Astigmatism	$\theta^2/2F_1$	$2(2F_1D/\lambda)^{1/2}$
Field curvature ^b	θ^2D/r_2	$2(r_2/\lambda)^{1/2}$

^a θ is the field angle, $F = f/D$ is the final focal ratio, $F_1 = f_1/D$ is the primary focal ratio, and r_2 is the radius of curvature of the secondary mirror. The terms for astigmatism and field curvature assume small back focal distance and large secondary magnification, which is typical for submillimeter telescopes. The Cassegrain and Gregory designs have the same coma and the same astigmatism

^b For the Cassegrain design, the focal surface is concave viewed from the secondary. For the Gregory, the focal surface is convex

and conic constant

$$b_2 = - \left(\frac{m_2 - 1}{m_2 + 1} \right)^2. \quad (7.4)$$

Sign conventions for (7.1–7.4) are shown in Table 7-2.

Table 7-3 shows aberrations for classical Cassegrain and Gregory telescopes. A submillimeter telescope typically has a primary focal ratio ≤ 0.7 , so field curvature and astigmatism are the largest aberrations. Field curvature can be corrected by a field-flattening lens, or by tiling the detectors on a faceted approximation to the focal surface, so astigmatism limits the FOV. It is difficult to make fast mirrors for optical wavelengths, so optical telescopes typically have a primary focal ratio in the range 1–2, and coma becomes more important. Third order coma can be eliminated by changing the conic constants of the primary and secondary to

$$b'_1 = -1 - \frac{2L}{d_1 m_2^3}$$

$$b'_2 = - \frac{2f}{d_1 (m_2 + 1)^3} - \left(\frac{m_2 - 1}{m_2 + 1} \right)^2, \quad (7.5)$$

resulting in the RC and aplanatic Gregory designs. The absence of coma and short overall length makes the RC form the choice for most optical and IR telescopes. Adjusting the conic constants of the primary and secondary does mean that the optical configuration is fixed, so it is not possible to change the final focal length by installing a new secondary, as is the case for the classical forms. The hyperboloidal RC primary also renders prime focus useless without a corrector. Since third-order coma does not limit the FOV in a telescope with a fast primary, all existing submillimeter telescopes are classical Cassegrain and Gregory designs. However, the development of large submillimeter detector arrays is pushing new telescopes to much wider FOV, which requires optical designs with good correction of higher order aberrations. As an example, the 25-m-diameter CCAT is a RC design with a shaped tertiary. This gives 0.9° FOV at $\lambda = 350 \mu\text{m}$, compared with 0.7° for a similar classical design with a shaped tertiary.

If the blockage due to the secondary and its support cannot be tolerated, an off-axis design must be used. This is often the case at longer wavelengths, where scattering from the telescope would exceed the loss through the atmosphere, or for observations of low surface brightness extended sources where false signals due to scattering and sidelobes would be a problem. The Gregory form is generally used for wide-field, off-axis designs because it gives a less offset primary and easy access to a pupil for chopping and scanning. Off-axis optical systems suffer from severe astigmatism and coma, but the secondary can be tilted to correct astigmatism (Dragone 1978, 1982; Hanany and Marrone 2002; Rusch et al. 1990), and the mirrors can be shaped to reduce coma, resulting in a FOV similar to that of an on-axis design with the same primary diameter and focal length. The secondary must be tilted to satisfy

$$m'_2 \tan i_1 = (m'_2 + 1) \tan i_2, \quad (7.6)$$

where m'_2 is the magnification of the tilted secondary and i_1 and i_2 are the angles of incidence of the principal ray at the primary and secondary. Examples of off-axis Gregory telescopes include AST/RO, SPT, ACT, WMAP, and Planck. With the exception of AST/RO, these are all millimeter-wave telescopes designed for CMB observations.

Most submillimeter instruments have a cold stop to control the illumination on the primary, so the entrance pupil is a little smaller than the primary mirror. In some cases, the stop can be at the secondary mirror, with the spillover falling on the cold sky, or on a cooled absorber (Padin et al. 2008a).

The main disadvantage of the Cassegrain and Gregory designs is the tight alignment tolerance. Most submillimeter telescopes provide active control of the secondary position to compensate gravitational deflection of the secondary support.

3.1.2 Plate Scale


In on-axis designs, the Cassegrain or Gregory focus is usually placed just behind the primary, partly to avoid blockage due to placing a large camera in the telescope beam and partly to give easy access to the camera. This configuration requires a telescope focal ratio greater than $\sim f/5$, which may lead to unreasonably large detectors. Submillimeter telescopes generally have a focal reducer to couple the telescope to the detectors.

For maximum aperture efficiency, a detector should just fill the central part of the Airy pattern delivered by the telescope, so the detector diameter must be $2F\lambda$, where F is the focal ratio at the detector. Arrays of $2F\lambda$ feedhorn-coupled detectors are often used for small cameras. The telescope FWHM beamwidth corresponds to $F\lambda$ at the detectors, so a detector array that Nyquist samples the sky must have $0.5F\lambda$ detectors. If the number of detectors is limited, e.g., by cost, $2F\lambda$ detectors give the highest mapping speed; otherwise, smaller pixels are better (Griffin et al. 2002). The typical size of an absorber-coupled submillimeter detector is 1 mm, so a $0.5F\lambda$ detector must be fed at $\sim f/3$. This usually requires a focal reducer, which can be a single off-axis ellipsoid for small FOV (Serabyn 1997). A focal reducer for larger FOV might use a pair of off-axis ellipsoids with an intermediate focus. Aberrations from the first off-axis ellipsoid can be at least partly compensated by the second (Serabyn 1995).

In an off-axis telescope, the camera can be placed in front of the primary without blocking the beam. In this case, the telescope focal ratio can be made small enough to directly feed the detectors, leading to a very simple, low-loss optical system (Padin et al. 2008a). The main

disadvantage of such an approach is that the camera must be mounted close to the secondary mirror, so access is difficult and the camera moves around as the telescope tracks and scans.

3.1.3 Other Telescope Designs

Cassegrain and Gregory designs dominate at submillimeter wavelengths, but other designs are sometimes used. The crossed Mizuguchi-Dragone telescope is an off-axis design with a large secondary (see  [Fig. 7-14](#)). It offers significantly wider FOV and better polarization performance compared with an off-axis Gregory of the same size (Dragone 1983; Tran et al. 2008). The crossed design has been proposed for CMB observations that do not require a large telescope. Refracting telescopes have been built for similar applications (Keating et al. 2003). Refracting optics are widely used inside submillimeter cameras and spectrometers, but dielectrics are quite lossy at submillimeter wavelengths, so thick lenses must be cooled. This severely limits the size of a refracting telescope.

3.1.4 Chopping and Scanning

Cassegrain telescopes are often equipped with a chopping secondary which allows a small field to be switched rapidly between the source and a reference position. Differencing the on and off source positions removes offsets and fluctuations in sky brightness (due to variations in the water vapor column) on timescales longer than the switching period. Chopping is an important technique for ground-based spectroscopy because sky brightness fluctuations typically dominate the noise. The throw for a chopping secondary is usually of order 10 beamwidths, limited by aberrations due to the tilt of the secondary. The chopping period must be a few times smaller than the wind crossing time, which is ~ 1 s for a 10-m telescope with a typical 10 ms^{-1} wind speed (Radford et al. 2008). Fast chopping with a large secondary is challenging, so chopping secondary mirrors are often made of very lightweight materials, e.g., CFRP or beryllium.

In an off-axis Gregory telescope, the pupil just after the secondary is a convenient, accessible location for a chopping mirror. In this case, the chopping mirror is a flat that can be tilted to generate a phase gradient across the pupil, which is equivalent to shifting the telescope beam on the sky. Chopping at a pupil allows a large chop throw without degrading the image quality over a wide FOV. Examples of telescopes with chopping flats include AST/RO and the Viper millimeter-wave telescope (Peterson et al. 2000). The development of large detector arrays has significantly reduced the requirements for chopping because sky brightness fluctuations are correlated across the detector array, so much of the sky signal can be removed by subtracting the average over the array (Jenness et al. 1998; Sayers et al. 2010).

For imaging observations, the telescope beam is usually scanned in a pattern that gives uniform coverage across the source, e.g., a raster or Lissajous pattern. Scanning fully samples the source, and allows the removal of offsets and sky brightness fluctuations that appear as low-frequency variations in the detector timestreams. During a scanning observation, it is usually sufficient for the telescope to follow the commanded path within about $1/2$ a beamwidth to maintain adequate coverage. The actual path is then accurately reconstructed after the observation using encoder readings and models of the telescope deformation due to acceleration. A scanning flat at a pupil can be used to scan the beam, and may be appropriate if very fast scanning is needed, but now it is often possible to scan the entire telescope at speeds of

order 1° s^{-1} . The SPT and ACT millimeter-wave CMB experiments, which are both Gregory designs, scan the entire telescope rather than a flat at a pupil. The major disadvantage of scanning the entire telescope is mechanical noise, which can cause problems for refrigerators and high-impedance detector readouts.

3.1.5 Scattering and Loss

Scattering and loss in the telescope degrade the aperture efficiency, increase the loading on the detectors, and may cause false signals. The main scattering contributions are blockage due to the secondary (typically 1% in an on-axis telescope), the secondary support (a few percent) (Cheng and Mangum 1998; Lamb and Olver 1986), gaps between mirror segments (usually $<1\%$), and surfaces of refractive elements inside cameras and spectrometers. Some of the scattering from the telescope can be directed to the sky, e.g., by using shaped secondary support legs (Lawrence et al. 1994; Moreira et al. 1996) and covering segment gaps with reflecting strips (Padin et al. 2008b), but some of the scattered light will be absorbed at ambient temperature.

On-axis telescope designs are generally adequate for ground-based observations at $\lambda < 1 \text{ mm}$ because the atmospheric transmission is worse than the telescope transmission. At longer wavelengths, the atmosphere is more transparent, so blockage is more important. This is one of the reasons millimeter-wave CMB experiments favor off-axis telescopes.

3.1.6 Optical Components

Primary mirrors for large telescopes must be segmented. Segments for submillimeter telescopes are made from machined Al (e.g., ALMA Vertex), electroformed Ni facesheets with an Al honeycomb core (e.g., ALMA Alcatel), Al facesheets with an Al honeycomb core (e.g., CSO), and aluminized CFRP with an Al honeycomb core (e.g., SMT). Conventional milling can achieve surface errors of $\sim 2 \mu\text{m}$ rms on a 0.5-m Al segment. The reflectivity of a machined Al mirror is $\sim 1 - 0.06 \times \exp(-\lambda/200 \mu\text{m})$ (Baars et al. 2006). Metal mirrors are too heavy for space telescopes, so CFRP facesheets with a CFRP core (e.g., Planck) or lightweighted SiC (e.g., Herschel) are used. These technologies are generally too expensive for ground-based telescopes.

The mass of the primary must be minimized because it is a strong driver for the overall mass and cost of a telescope. Machined Al segments can achieve an areal density of $\sim 20 \text{ kg m}^{-2}$, which is about the same as the areal density of the entire Herschel primary. The choice of segment size is a trade-off between manufacturing errors and thermal deformations, which increase with segment size, vs. areal density, complexity, and noise for active control, all of which decrease with segment size. Most submillimeter telescopes have 1–2-m segments.

Plastics are widely used for submillimeter lenses and windows, with machined grooves (Goldsmith 1998) or plastic foam anti-reflection (AR) coatings (Savini and Hargrave 2010). HDPE and UHMWPE (refractive index $n = 1.5$, loss tangent $\tan \delta \sim 5 \times 10^{-4}$ at $\lambda = 350 \mu\text{m}$) are popular because they have low loss in the submillimeter, but other plastics, e.g., teflon, nylon, and polystyrene, are used, particularly at longer wavelengths (Lamb 1996). High-resistivity Si ($n = 3.42$, $\tan \delta \sim 10^{-5}$ at $\lambda = 1 \text{ mm}$) has lower loss, and the high-refractive index generally gives better image quality, but AR coating is more difficult. AR coatings can be etched on small Si optics and polyimide (Fowler et al. 2007) layers have been used on larger Si lenses.

3.2 Structure and Mechanics

Structural and mechanical systems support the optical surfaces as the telescope scans and tracks a source. The quality of the wavefront delivered to the detectors is generally determined by the telescope structure, so this is a critical aspect of the telescope design. The following sections describe the key telescope systems and some simple models that illustrate typical performance at submillimeter wavelengths.

3.2.1 Pointing and Wavefront Errors

The telescope mount usually determines the pointing error, while the primary mirror support determines the mirror surface error and hence higher-order wavefront errors. All these errors are related in that a pointing error is just a tilt, which is the lowest-order wavefront error. It is usual to carry pointing and surface error as separate terms in the telescope error budget because the pointing error can easily be measured and corrected by observing a nearby, bright, point source. At submillimeter wavelengths, atmospheric seeing does not dominate the wavefront error under good observing conditions, so the telescope sees a diffraction-limited image (with resolution $1.22\lambda/D$) that moves around slowly as the telescope pointing error varies (typically $1\text{--}2''$ on timescales of hours).

3.2.2 Primary Mirror Support

The primary mirror is the most difficult part of any reflecting telescope. The key design consideration is maintaining the mirror surface accuracy in the presence of gravity, temperature variations, wind, solar illumination, and acceleration due to scanning. Gravitational and thermal deformations are the biggest problems for ground-based telescopes.

Telescope primary mirrors are essentially plate structures, so a simple model of the sag of a plate supported around its edge can provide some insight. The p-p deflection of a simply supported plate is (Timoshenko and Woinowsky-Krieger 1959)

$$\delta \approx \frac{qD^4}{24E\eta h^3}, \quad (7.7)$$

where q is the pressure on the plate, h is the plate thickness, E is Young's modulus for the plate material, and η is the filling factor. The pressure on the plate due to its own weight and the weight of the segments is

$$q \sim 2gh\rho\eta, \quad (7.8)$$

where g is the acceleration due to gravity and ρ is the density of the plate material. The factor 2 in (7.8) is a somewhat pessimistic estimate of the areal density of the segments. The thickness of a submillimeter primary mirror is typically similar to the depth of the primary surface, which is $D/8$ for a focal ratio of 0.5, so the rms gravitational deflection is

$$\sigma \approx \frac{\delta}{3} \sim \frac{2g\rho D^2}{E}, \quad (7.9)$$

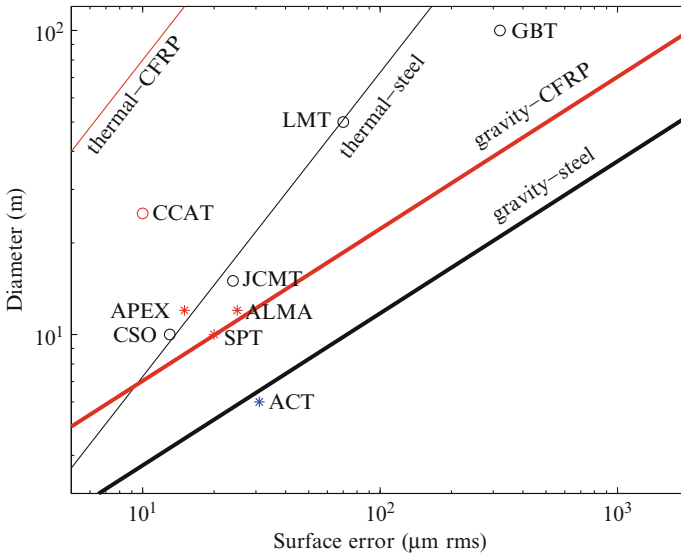


Fig. 7-15

Diameter vs. surface error for passive (*) and active (o) primary mirror supports made of steel (black), CFRP (red), and aluminum (blue). Points for CCAT and LMT are predictions. Materials properties are given in Table 7-4

where the rms is taken to be one third of the p-p. Temperature variations across the plate change its thickness, resulting in rms surface error

$$\sigma_T \approx \alpha T (D/8), \quad (7.10)$$

where T is the rms temperature variation and α is the coefficient of thermal expansion of the plate material. Equations 7.9 and 7.10 are plotted in Fig. 7-15 for different materials and 1-K-rms temperature variation, which is typical for a submillimeter telescope (Greve and Bremer 2010). The primary diameter vs. surface error plot is often called a von Hoerner plot after its inventor (von Hoerner 1967). The plot is useful because it indicates which materials must be used to achieve a given performance with a passive support structure, or alternatively when active control of the optical surfaces is needed. Passive supports are often designed to deform under gravity in a way that maintains a paraboloidal surface, but with an elevation-dependent focal length. These homologous designs achieve gravitational deflections a few times smaller than predicted by 7.9. The CSO uses a combination of homology and slow active control to beat the limits for gravitational deflection of its steel structure, so the performance is limited by thermal deformations. The GBT and LMT are examples of much larger, lower-frequency, telescopes with steel structures and fast active surfaces (Baars 2007). ALMA uses a CFRP structure to achieve the required performance with a completely passive support. ACT is made of Al, so its performance is limited by thermal deformations.

For a given surface error in **Fig. 7-15**, the Strehl ratio is

$$S = \exp - \left(\frac{4\pi\sigma}{\lambda} \right)^2, \quad (7.11)$$

which is the Ruze formula familiar to designers of microwave telescopes (Ruze 1966). High Strehl ratio places stringent demands on the surface error. For example, $S > 0.9$, corresponding to <20% increase in integration time compared with an ideal telescope, requires $\sigma < \lambda/39$, which is challenging at short wavelengths.

During a scanning observation, acceleration of the telescope causes the primary to deflect, resulting in pointing errors that can degrade the image resolution. The effect is small for existing submillimeter telescopes, which have large beamwidths and slow drives, but it will be important for future, large, submillimeter telescopes that scan quickly. The pointing error can be calculated in terms of the gravitational deflection or natural frequency, leading to some useful expressions that capture the dynamic performance of the primary. The stiffness of the primary is

$$k = \frac{mg}{\delta}, \quad (7.12)$$

where m is the mass of the primary, so the natural frequency is

$$\omega = \sqrt{\frac{k}{m}} = \sqrt{\frac{g}{\delta}} \sim \frac{1}{D} \sqrt{\frac{E}{6\rho}}. \quad (7.13)$$

The deflection of the primary due to angular acceleration Ω is

$$\delta_a \approx \Omega (D/4) \delta / g \approx \Omega (D/4) / \omega^2, \quad (7.14)$$

and the corresponding pointing error is

$$\theta_a \approx \delta_a / (D/4) \approx \Omega / \omega^2. \quad (7.15)$$

A 25-m primary made of steel should achieve ~14-Hz natural frequency, so at 1° s^{-2} acceleration, the pointing error should be $0.47''$, which is 1/7 of the beamwidth at $\lambda = 350 \mu\text{m}$, essentially the entire pointing budget. If the natural frequency were only 7 Hz, which is probably more realistic, the pointing error would be $>1/2$ the beamwidth, which is enough to severely degrade the image resolution.

Submillimeter telescopes often have a simple enclosure to protect the telescope from wind buffeting, solar heating, and severe weather (see **Table 7-1**). Wind protection is important for telescopes with light structures because these designs tend to have low stiffness, even though the stiffness to mass ratio is high enough to give small gravitational deformation. The rms wind-induced deformation of the primary is

$$\sigma_w \approx \sigma \frac{q_w}{q}, \quad (7.16)$$

where

$$q_w = \frac{1}{2} \rho_{\text{air}} v^2 \quad (7.17)$$

is the wind pressure, $\rho_{\text{air}} \approx 1 \text{ kg m}^{-3}$ is the density of air, and v is the wind speed. Combining **(7.16)**, **(7.13)**, **(7.9)**, and **(7.8)** yields

$$\sigma_w \sim \frac{1}{2} \rho_{\text{air}} v^2 \frac{8D}{E\eta} \sim \frac{1}{2} \rho_{\text{air}} v^2 \frac{1}{\omega^2 D \rho \eta}. \quad (7.18)$$

The outside wind speed at a submillimeter site might be 10 ms^{-1} , so $q_w \approx 50 \text{ Pa}$. For a 25 m diameter steel primary with 7 Hz natural frequency and 1% filling factor, this wind pressure will cause $\sim 13 \mu\text{m}$ rms deformation, which is the entire surface error budget for a $\lambda = 350 \mu\text{m}$ telescope and an order of magnitude larger than the fraction of the budget that might typically be allocated to wind-induced deformation. If this primary were exposed, its stiffness would have to be increased by an order of magnitude. Stiffness scales roughly with filling factor, so the cost of the primary support would also increase by an order of magnitude. However, wind pressure scales with the square of the wind speed, so even a simple enclosure can make the wind pressure negligible. This approach can be much less expensive than making the structure stiff enough to resist outside wind forces.

3.2.3 Mirror Control

Most existing submillimeter telescopes have passive primary mirror support systems, which can achieve $\sim 15\text{-}\mu\text{m}$ -rms surface accuracy on a $\sim 10\text{-m}$ telescope. The CSO is an example of a submillimeter telescope with an active primary. Its segments are attached to a steel spaceframe truss with $\sim 100\text{-mm}$ -long steel rods that can be heated to give $\sim 10 \mu\text{m}$ of length adjustment. This allows slow active control which has improved the surface error from 22 to $\sim 13 \mu\text{m}$ rms (Leong et al. 2006; Woody et al. 1998). Future, larger telescopes will have larger gravitational deformations and will require fast active control.

The simplest approach to active control is open loop, based on look up tables for gravity and temperature. The GBT, LMT, JCMT, and CSO use this type of control. Open loop control requires a mirror support with good repeatability. In practice, open loop control is limited by the thermal stability of the mirror support because it is difficult to generate a thermal model that can predict deformations on small spatial scales.

Closed loop control is used on segmented optical telescopes and may be needed for future, large, submillimeter telescopes. The usual approach is to measure the relative piston and tilt of neighboring segments using edge sensors. These can be capacitive displacement sensors, as in the Keck telescopes (Nelson et al. 1985), inductive sensors, as in HET (Booth et al. 2003) and SALT (Menzies et al. 2010), or optical sensors, as proposed for CCAT (Woody 2011). Edge sensors have good sensitivity to deformations on the scale of a segment, but the sensitivity to low-order modes of the primary is poor because many sensor readings must be combined to measure a low-order deformation. This is not a big issue for an optical telescope, because a wavefront measurement on a bright star can quickly constrain the low-order modes, but it is a serious concern for a submillimeter telescope, where submillimeter wavefront measurements are slow and a suitable source is often not available. If the noise for a single sensor is e_s (in units of length) and the noise is uncorrelated between sensors, the noise for a measurement of the lowest-order mode of the primary is

$$e_u \sim e_s \sqrt{D/d}, \quad (7.19)$$

where d is the size of a segment and D/d is the number of sensors across the mode. The factor $\sqrt{D/d}$ in (7.19) is often called the error multiplier. For an optical telescope, the sensor noise is mainly electrical noise, but a submillimeter telescope is made of materials with higher and less uniform coefficient of thermal expansion, so thermal deformation of the sensor mounts dominates. A submillimeter mirror segment might have $e_s \sim 1 \mu\text{m}$ rms, so sensor noise is a serious issue for a large telescope. The situation is even worse for deformations that are correlated

between segments, e.g., gravitational deformation and cupping of the segments due to a thermal gradient through the segments. In this case, the noise for a measurement of the lowest-order mode is

$$e_c \sim e_s D/d. \quad (7.20)$$

In principle, an edge sensor system with enough sensors can measure cupping of the segments, so it may be possible to correct the effect of segment deformations in a closed loop control system.

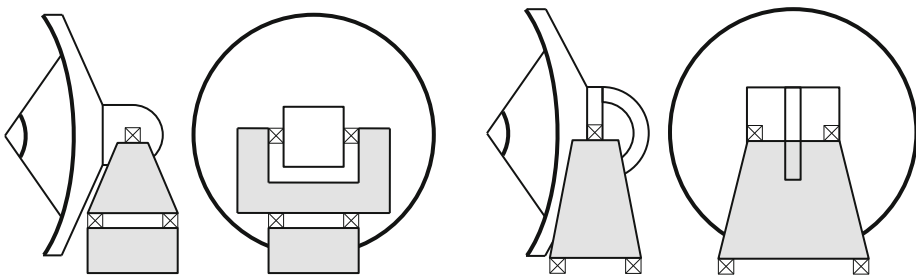
3.2.4 Telescope Mount

The telescope mount usually has little impact on the surface accuracy of the primary, but it does set the pointing performance. Most modern telescopes have an elevation over azimuth mount to minimize the mass and cost of the mount. There are many design variations, but elevation over azimuth mounts fall into two broad groups: a fork on a pedestal, or an alidade on a track (see [Fig. 7-16](#)). Mount structures are generally made of steel, to reduce the cost, and the filling factor for the structure is typically $\sim 1\%$ for box and spaceframe mount designs. The mounts in [Fig. 7-16](#) look roughly like beams $\sim D/2$ high $\times D/2$ wide (along the elevation axis) $\times D/4$ deep (fore-aft), so a simple beam model can provide a useful estimate of the performance. The stiffness of the beam for end loading is (Olberg et al. 1992)

$$k' \sim \frac{E' \eta' D}{64}, \quad (7.21)$$

where E' is Young's modulus for the mount material and η' is the filling factor. The primary requires a counterweight with mass similar to that of the primary, and an elevation axle that might be twice the mass of the primary, so the mass of the tipping structure is $\sim 4\times$ the mass of the primary. The mass of the mount is usually similar to the mass of the tipping structure, so the total mass of telescope is

$$m' \sim 16\pi (D/2)^2 (D/8) \eta \rho, \quad (7.22)$$



■ Fig. 7-16

Fork on pedestal (*left*) and alidade on track (*right*) telescope mounts. Crosses indicate bearings. Some fork mounts are inverted, with the fork arms attached to the primary rather than to the pedestal

where the mass of the primary is given by (7.8). The natural frequency of the telescope is then

$$\omega' \sim \sqrt{\frac{k'}{m'}} \sim \frac{1}{10D} \sqrt{\frac{E'\eta'}{\rho\eta}}. \quad (7.23)$$

For a steel structure, $\omega' / (2\pi) \sim 100/D$, where D is in meters. This is consistent with measured natural frequencies for radio telescopes in the 10–30-m-diameter range (Gawronski 2005).

Equation 7.23 is a rough but useful guide to the dynamic performance of the structure, and it immediately gives an estimate of the fast motion capabilities of the telescope because the drive control loop bandwidth must be a factor 4–10 smaller than the natural frequency. Thus, a 25-m steel telescope should achieve a drive bandwidth of 0.4–1 Hz, which would allow position switching at ~ 0.1 Hz. A CFRP primary would improve the drive bandwidth by roughly a factor 2.

A simple model can also give us a useful estimate of pointing errors due to temperature variations in the mount. For a massive structure like the mount, thermal deformations are mainly driven by diurnal temperature variations, while for lighter, open structures like the primary, spatial variations in air temperature are more important. The thermal time constant of the mount is

$$\tau = C/G, \quad (7.24)$$

where C is the specific heat and G is the thermal conductance. For pointing errors, we are concerned with temperature variations across the mount. Breaking the upper part of the mount into three slices, one for each elevation bearing and one in the center, gives

$$G \sim \beta (D/4)^2 \eta' / (D/6), \quad (7.25)$$

where β is the thermal conductivity of the mount material. The specific heat of one of the slices is

$$C \sim c (D/4)^2 (D/6) \eta' \rho', \quad (7.26)$$

where c and ρ' are the specific heat capacity and density of the mount material. The thermal time constant is then

$$\tau \sim \frac{c\rho'}{\beta} \times \frac{D^2}{36}. \quad (7.27)$$

For a steel mount, $\beta/c\rho' = 1.4 \times 10^{-5} \text{ m}^2 \text{ s}^{-1}$ (see Table 7-4), so $\tau \sim D^2/2 \text{ h}$, where D is in meters. If the mount acts as a single-pole, low-pass, thermal filter, the mount temperature is

Table 7-4
Material properties (Wilson 1999)

Property	Symbol	Units	Aluminum	Steel	CFRP
Density	ρ	kg m^{-3}	2,600	7,850	1,550
Young's modulus	E	Pa	7×10^{10}	2.1×10^{11}	1.5×10^{11}
Stiffness	E/ρ	J kg^{-1}	2.7×10^7	2.7×10^7	9.7×10^7
Coefficient of thermal expansion	α	K^{-1}	2.2×10^{-5}	1.1×10^{-5}	2×10^{-7} to 10^{-6}
Specific heat capacity	c	$\text{J kg}^{-1} \text{K}^{-1}$	890	460	
Thermal conductivity	β	$\text{W m}^{-1} \text{K}^{-1}$	160	49	
Thermal diffusivity	$\beta/(c\rho)$	$\text{m}^2 \text{s}^{-1}$	6.9×10^{-5}	1.4×10^{-5}	

$$T(t) \sim (\Delta T/\tau) \sin(2\pi t), \quad (7.28)$$

where ΔT is the amplitude of diurnal variations in air temperature. Temperature gradients across the mount can be as large as $T(t)$, in which case the pointing error is

$$\theta_T \sim \alpha' T(t), \quad (7.29)$$

where α' is the coefficient of thermal expansion of the mount material. Pointing errors are usually measured by observing a nearby bright point source, and the measurements are made often enough to keep the pointing error smaller than about 1/10 of the beamwidth. The time between pointing measurements is then

$$t' < (1/10) (1.22\lambda/D) / \dot{\theta}_T \sim \lambda D / (100\alpha' \Delta T) \text{ h}, \quad (7.30)$$

where D and λ are in meters. ΔT is typically 5 K at high desert sites (Radford et al. 2008), so for a 10-m steel telescope at $\lambda = 350 \mu\text{m}$, $t' < 3/4$ h. Pointing measurements typically take a few minutes, so a measurement roughly every hour does not significantly impact the observing efficiency. Since the thermal time constant scales with D^2 , larger telescopes have better mount thermal performance, despite the smaller beamwidth.

Telescope mounts are equipped with encoders to measure the position of the axes. Most modern telescopes use optical tape or disk encoders. A tiltmeter is often mounted on the azimuth axis to measure changes in the overall tilt of the structure. Some mounts include metrology systems to directly measure thermal deformations of the yoke, and some use temperature sensors to predict thermal deformations. All existing submillimeter telescope mounts have rolling element bearings, but hydrostatic bearings will likely be used on future, large, submillimeter telescopes. The drives are typically brushless DC motors with gearboxes driving a ring or sector gear, but some telescopes (e.g., ALMA, Mangum et al. 2006) have direct drives.

3.3 Alignment

The telescope optics must be aligned when the telescope is built, and alignment must be maintained during observations. This is a difficult problem. Some of the techniques that have been developed for aligning submillimeter telescopes are described below. These fall into two groups: direct measurements of the mirror surfaces and measurements of the wavefront delivered to a detector. In general, telescope alignment starts with direct measurements and then moves to wavefront measurements when the errors are much smaller than the observing wavelength.

Initial alignment of submillimeter telescope surfaces used to be done with a theodolite and tape, which can achieve a few $\times 100\text{-}\mu\text{m}$ -rms surface errors on a 10-m telescope, but photogrammetry and laser trackers are now widely used. Photogrammetry involves taking photographs of reflective targets on the primary from different angles. A calibrated ruler is included somewhere in the photographs to give the overall scale. Photogrammetry can measure features to about one part in 10^5 , i.e., a few $\times 10 \mu\text{m}$ rms for a 10-m primary. This technique was used for initial alignment of ALMA, and SPT. Laser trackers use precise measurements of the angle of a laser beam reflected from a retro-reflector on the surface being measured, combined with interferometry to measure the distance to the surface. The retro reflector is usually a corner cube mounted in a sphere that can be moved around on the surface. A laser tracker can measure features to a few $\times 10 \mu\text{m}$ rms on a 10-m primary in the field, but much more accurate results can be achieved

under controlled conditions (Zobrist et al. 2009). ACT was aligned using a laser tracker (Hincks et al. 2008).

Following initial alignment, millimeter-wave holography (Bennett et al. 1976; Scott and Ryle 1977) is often used to measure and correct the surface. Holography measurements can easily achieve an accuracy of a few microns. The technique measures the far field response of a telescope, which is the Fourier Transform of the complex illumination pattern on the primary. The phase of the illumination pattern corresponds to the wavefront error. A measurement of the far field response requires an interferometer in which the telescope under test is scanned across a source while the reference telescope is fixed on the source. A bright celestial source, e.g., a planet, can be used to measure the wavefront over a range of elevations. This type of measurement typically requires a large reference telescope, so it is easy for telescopes in an array, but rarely practical for a single telescope. An artificial source on a tower can be used (Baars et al. 2007), in which case the reference telescope might be just a small horn, but this measurement yields the near field response. The far field response can be calculated if the relative position of the source and telescope is known. For some near field holography measurements, the reference horn is mounted on the back of the secondary of the telescope being measured. In this case, the horn also moves, but it is small, so its response does not change much. The main problem with measurements that use an artificial source on a tower is that the surface can be measured at only one, low elevation.

Out of focus holography (Morris et al. 1991; Nikolic et al. 2007a) can be used to measure low-order wavefront errors because they change the shape of the PSF. The technique involves measuring the PSF at and on either side of focus, e.g., by moving the secondary while observing a bright point source with a camera. Model fits to the PSF yield the first few modes of the wavefront. This is useful because the largest errors are typically in the lowest-order modes. Out of focus holography is routinely used to measure and correct the 100-m diameter GBT (Nikolic et al. 2007b).

The CSO uses a shearing interferometer (Serabyn et al. 1991) to measure wavefront errors. This is another way of measuring the far field response of the telescope, but instead of using a reference telescope, as in holography, light from the core of an image of a point source provides the reference. The beam from the telescope is split and then recombined at a focus so that one point source image (the signal) can be moved relative the other (the reference). A path length modulator is included in the reference arm so that a total power detector can measure the complex field vs. frequency where the reference and signal images overlap. The signal image is stepped across the reference until the response function has been sampled to large enough radius to give the desired spatial resolution on the primary.

Phase contrast interferometry (Dicke 1975; Malacara 1992; Serabyn and Wallace 2010) is another promising technique for submillimeter wavefront measurements. In this case, a camera is used to inspect the pupil while the average phase of the illumination is changed. A $\pi/2$ phase shift converts small phase variations into brightness variations, so a measurement of the brightness over the pupil gives the wavefront error. The phase shift is done by modulating the path length at the core of a point source image, either using phase plates or a small moving mirror in the center of a larger mirror. CCAT will use phase contrast interferometry to set and maintain its optics. A measurement of the 25-m surface with 0.5-m spatial resolution and a few μm rms accuracy will take ~ 10 s on Mars to ~ 2 h on Uranus, depending on the wavelength (Serabyn 2006).

References

- Baars, J. W. M. 2007, *The Paraboloidal Reflector Antenna in Radio Astronomy and Communication* (New York: Springer)
- Baars, J. W. M., et al. 1994, *Proc. IEEE*, 82, 687–696
- Baars, J. W. M., et al. 2006, ALMA Memo 566 NRAO, Charlottesville VA
- Baars, J. W. M., et al. 2007, *IEEE Antennas Propag. Mag.*, 49, 24–41
- Bennett, J. C., et al. 1976, *IEEE Trans. Antennas Propag.* 24, 295–303
- Blake, G. A., Keene, J., & Phillips, T. G. 1985, Chlorine in dense interstellar clouds – the abundance of HCl in OMC-1. *ApJ*, 295, 501
- Blake, G. A., Sutton, E. C., Masson, C. R., & Phillips, T. G. 1987, Molecular abundances in OMC-1 – the chemical composition of interstellar molecular clouds and the influence of massive star formation. *ApJ*, 315, 621
- Blundell, R. 2007, in *IEEE/MTT-S International Microwave Symposium* (New York: IEEE), 1857–1860
- Booth, J. A., et al. 2003, *Proc. SPIE*, 4837, 919–933
- Carlstrom, J. E., et al. 2011, *PASP*, 123, 568–581
- Caves, C. 1982, Quantum limits on noise in linear amplifiers. *Phys. Rev. D*, 26, 1817
- Cheng, J., & Mangum, J. G. 1998, MMA Memo, Vol. 197 (Tucson: NRAO, Charlottesville VA)
- Dayem, A. H., & Martin, R. J. 1962, Quantum interaction of microwave radiation with tunneling between superconductors. *Phys. Rev. Lett.*, 8, 246
- Dicke, R. H. 1975, *ApJ*, 198, 605–615
- Dolan, G. J., Phillips, T. G., & Woody, D. P. 1979, Low-noise 115 GHz mixing in superconducting oxide-barrier tunnel junctions. *Appl. Phys. Lett.*, 34, 347
- Dragone, C. 1978, *Bell Syst. Tech. J.*, 57, 2663–2684
- Dragone, C. 1982, *IEEE Trans. Antennas Propag.*, 30, 331–339
- Dragone, C. 1983, *IEEE Trans. Antennas Propag.*, 31, 764–775
- Fowler, J. W., et al. 2007, *Appl. Opt.*, 46, 3444–3454
- Gawronski, W. 2005, in *American Control Conference*. IEEE, FrA11.2
- Gershenson, E., Gol'tsman, G., Gogidze, I. G., Gusev, Y. P., Elant'ev, A. I., Karasik, B. S., & Semenov, A. 1990, Millimeter and submillimeter range mixer based on electronic heating of superconducting films in the resistive state. *Sov. Phys. Supercond.*, 3, 1582
- Goldsmith, P. F. 1998, *Quasioptical Systems* (Piscataway: IEEE). **Chap. 5**
- Greve, A., & Bremer, M. 2010, *Thermal Design and Thermal Behaviour of Radio Telescopes and their Enclosures* (Berlin: Springer). Chap. 9
- Griffin, M. W., Bock, J. J., & Gear, W. K. 2002, *Appl. Opt.*, 41, 4666–4670
- Guilloteau, S., et al. 1992, *A&A*, 262, 624–633
- Gusten, R., et al. 2006, *A&A*, 454, L13–L16
- Hanany, S., & Marrone, D. P. 2002, *Appl. Opt.*, 41, 4666–4670
- Hincks, A. D., et al. 2008, *Proc. SPIE*, 7020, 70201P
- Jaffe, D. T., Harris, A. I., Silber, M., Genzel, R., & Betz, A. L. 1985, *ApJ*, 290, L59
- Jenness, T., Lightfoot, J. F., & Holland, W. S. 1998, *Proc. SPIE*, 3357, 548–558
- Keating, B. G., et al. 2003, *Proc. SPIE*, 4843, 284–295
- Keene, J., Blake, G. A., & Phillips, T. G. 1983, First detection of the ground state $J_K = 1_0 \rightarrow 0_0$ submillimeter transition of interstellar ammonia. *ApJ*, 271, L27
- Kooi, J. W., Chan, M., Phillips, T. G., Bumble, B., & LeDuc, H. G. 1992, *IEEE Trans. Microw. Theory Technol.*, 40, 812
- Korsch, D. 1991, *Reflective Optics* (San Diego: Academic Press)
- Kramer, C., et al. 1998, *Proc. SPIE*, 3357, 711–720
- Lamb, J. W. 1996, *Int. J. Infrared Millim. Waves*, 17, 1997–2034
- Lamb, J. W., & Olver, A. D. 1986, *Proc. IEE-H*, 133, 43–49
- Lawrence, C. R., Herbig, T., & Readhead, A. C. S. 1994, *Proc. IEEE*, 82, 763–767
- Leighton, R. B. 1978, A 10 meter telescope for millimeter and submillimeter astronomy. Final Technical Report for NSF Grant 73-04908
- Leong, M., et al. 2006, *Proc. SPIE*, 6275, 62750P
- Low, F. J. 1961, Low-temperature germanium bolometer. *JOSA*, 51(11), 1300–1304
- Malacara, D. 1992, *Optical Shop Testing* (2nd ed.; New York: Wiley). Chap. 3
- Mangum, J. G., et al. 2006, *PASP*, 118, 1257–1301
- McGrath, W. R. et al. 1997, Superconductive hot electron mixers with ultra-wide RF bandwidth for heterodyne receiver applications up to 3 THz, in *Proceedings of the ESA Symposium 'The Far Infrared and Submillimetre Universe'*, Grenoble, France, ESA SP-401
- Menzies, J., et al. 2010, *Proc. SPIE*, 7739, 77390X
- Moreira, F. J. S., Prata, A., Jr., & Thorburn, M. A. 1996, *IEEE Trans. Antennas Propag.*, 44, 492–499
- Morris, D., Davis, J. H., & Mayer, C. E. 1991, *Proc. IEE-H*, 138, 243–247

- Morris, D., et al. 2009, *IET Microw. Antennas Propag.* 3, 99–108
- Nelson, J. E., Mast, T. S., & Faber, S. M. 1985, *Keck Obs. Rep.* 90, *The Design of the Keck Observatory and Telescope* (Berkeley: W. M. Keck Observatory, Kamuela HI)
- Neufeld, D. A., et al. 2010, Strong absorption by interstellar hydrogen fluoride: Herschel/HIFI observations of the sight-line to G10.6–0.4 (W31C). *A&A*, 518, L108
- Nikolic, B., Hills, R. E., & Richer, J. S. 2007a, *A&A*, 465, 679–683
- Nikolic, B., et al. 2007b, *A&A*, 465, 685–693
- Olberg, E., et al. 1992, *Machinery's Handbook* (24th ed.; New York: Industrial Press), 226
- Padin, S. et al. 2008a, *Appl. Opt.*, 47, 4418–4428
- Padin, S., et al. 2008b, *Electron. Lett.*, 44, 950–952
- Page, L., et al. 2003, *ApJ*, 585, 566–586
- Pearson, J. C., Guesten, R., Klein, & T., Whyborn, N. D. 2000, The local oscillator system for FIRST (HIFI), in *Proceedings of the SPIE 4013, UV, Optical, and IR Space Telescopes and Instruments*, ed. J. B. Breckinridge, & P. Jacobsen, Munich, Germany (Bellingham: SPIE), 264
- Peterson, J. B., et al. 2000, *ApJ*, 532, L83–L86
- Phillips, T. G. 1988, *Techniques of submillimeter astronomy*, in *Millimetre and Submillimetre Astronomy*, ed. R. D. Wolstencroft, & W. B. Burton (Dordrecht/Boston: Kluwer), 1–25.
- Phillips, T. G., & Huggins, P. J. 1981, Abundance of atomic carbon (C I) in dense interstellar clouds. *ApJ*, 251, 533–540
- Phillips, T. G., & Jefferts, K. B. 1973, A cryogenic bolometer heterodyne receiver for millimeter wave astronomy. *RSci*, 44, 1009
- Phillips, T. G., & Keene, J. 1992, *Submillimeter astronomy*. *IEEE Proc.*, 80, 1662
- Phillips, T. G., & Woody, D. P. 1982, *Millimeter- and submillimeter-wave receivers*. *ARA&A*, 20, 285
- Phillips, T. G., Neugebauer, G., Werner, M. W., & Huggins, P. J. 1977, Detection of submillimeter (870-micron) CO emission from the Orion molecular cloud. *ApJ*, 217, L161
- Phillips, T. G., Kwan, J. Y., & Huggins, P. J. 1980a, Detection of submillimeter lines of CO (0.65 mm) and H₂O (0.79 mm), in *IAU Symp.* 87, *Interstellar Molecules* (Dordrecht: Reidel), 21
- Phillips, T. G., Huggins, P. J., Kuiper, T. B. H., & Miller, R. E. 1980b, Detection of the 610 μ m (492 GHz) line of interstellar atomic carbon. *ApJ*, 238, L103
- Pilbratt, G. L., et al. 2010, *A&A*, 518, L1
- Prober, D. I. 1993, Superconducting terahertz mixer using a transition-edge microbolometer. *Appl. Phys. Lett.*, 62, 2119
- Radford, S., et al. 2008, *Proc. SPIE*, 7012, 70121Z
- Richards, P. L., Shen, T. M., Harris, R. E., & Lloyd, F. L. 1979, A quasiparticle heterodyne mixing in SIS tunnel junctions. *Appl. Phys. Lett.*, 34, 345
- Rowan-Robinson, M., Broadhurst, T., Lawrence, A., McMahon, R. G., Lonsdale, C. J., Oliver, S. J., Taylor, A. N., Hacking, P., Conrow, T., Saunders, W. S., Ellis, R. S., Efstathiou, G. P., & Condon, J. J. 1991, A high-redshift IRAS galaxy with huge luminosity – hidden quasar or protogalaxy? *Nature*, 351, 719
- Rusch, W. V. T. et al. 1990, *IEEE Trans. Antennas Propag.*, 38, 1141–1149
- Russell, R. W., Melnick, G., Gull, G. E., & Harwit, M. 1980, Detection of the 157 micron (1910 GHz) [C II] emission line from the interstellar gas complexes NGC 2024 and M42. *ApJ*, 240, L99
- Ruze, J. 1966, *Proc. IEEE*, 54, 633–640
- Saito, M. 2011, in *Proceedings of the General Assembly and Scientific Symposium 2011 XXXth URSI, Istanbul*, J10.6
- Savini, G., & Hargrave, P. C. 2010, *Proceedings of the 35th International Conference on Infrared, Millimeter and Terahertz Waves* (Piscataway: IEEE), 1–2
- Sayers, J., et al. 2010, *ApJ*, 708, 1674–1691
- Schroeder, D. R. 2000, *Astronomical Optics* (2nd ed.; San Diego: Academic). Chap. 6
- Scott, P. F., & Ryle, M. 1977, *Mon. Not. R. Astron. Soc.* 178, 539–545
- Serabyn, E. 1995, Wide-field imaging optics for submm arrays, *ASP Conference Series* 75, 74–81, *Astronomical Society of the Pacific*, Orem UT
- Serabyn, E. 1997, *Int. J. Infrared Millim. Waves*, 18, 273–284
- Serabyn, E. 2006, *Proc. SPIE*, 6275, 62750Z
- Serabyn, E., & Wallace, J. K. 2010, *Proc. SPIE*, 7741, 77410U
- Serabyn, E., Phillips, T. G., & Masson, C. R. 1991, *Appl. Opt.*, 30, 1227–1241
- Stacey, G. J., Lugten, J. B., & Genzel, R. 1987, Detection of interstellar CH in the far-infrared. *ApJ*, 313, 859
- Stark, A. A., et al. 2001, *PASP*, 113, 567–585
- Storey, J., Watson, D., & Townes, C. 1981, Detection of interstellar OH in the far-infrared. *ApJ*, 244, L27
- Swetz, D. S., et al. 2011, *ApJ Suppl.*, 194, 41–
- Tauber, J. A., et al. 2010, *A&A*, 520, A2
- Timoshenko, S., & Woinowsky-Krieger, S. 1959, *Theory of Plates and Shells* (2nd ed.; New York: McGraw-Hill), 57
- Tran, H., et al. 2008, *Appl. Opt.*, 47, 103–109
- Tucker, J. R. 1979, Quantum limited detection in tunnel junction mixers. *IEEE J. Quantum Electron.*, 15, 1234

- von Hoerner, S. 1967, *AJ*, 72, 35–47
- Wengler, M. J., & Woody, D. P. 1987, Quantum noise in heterodyne detection. *IEEE J. Quantum Electron.*, 23, 613
- Wilson, R. N. 1996, *Reflecting Telescope Optics I* (Berlin: Springer)
- Wilson, R. N. 1999, *Reflecting Telescope Optics II* (Berlin: Springer)
- Woody, D. P. 2009, *Submillimeter Astrophysics and Technology: A Symposium Honoring Thomas G. Phillips*, ASP Conference Series, 417, 3
- Woody, D. P. 2011, in *Proceedings of the General Assembly and Scientific Symposium 2011 XXXth URSI, Istanbul*, J10.7
- Woody, D., Serabyn, E., & Shinckel, A. 1998, *Proc. SPIE*, 3357, 474–485
- Zobrist, T. L., et al. 2009, *Proc. SPIE*, 7426, 742613

8 Radio Telescopes

Ron Ekers¹ · Thomas L. Wilson²

¹Australia Telescope National Facility, CSIRO Astronomy and Space Science, Epping, NSW, Australia

²Naval Research Laboratory, Washington, DC, USA

1	<i>Introduction</i>	318
2	<i>History</i>	319
2.1	Early History	319
2.2	Evolution of Radio Telescope Sensitivity	319
2.2.1	Exponential Growth in Science	319
2.2.2	Livingston Curve	320
2.3	The Development of the Aperture Synthesis Radio Telescope	321
2.3.1	Australian Group	321
2.3.2	Cambridge Group	321
2.3.3	The Beginnings of Aperture Synthesis	321
2.3.4	Earth Rotation Synthesis	322
3	<i>Radio Astronomy Fundamentals</i>	322
3.1	Radiative Transfer and Black Body Radiation	322
3.2	The Nyquist Theorem and Noise Temperature	323
3.3	Overview of Intensity, Flux Density, and Main Beam Brightness Temperature	324
3.4	Polarization	324
3.5	Sensitivity	325
4	<i>Antennas</i>	325
4.1	The Hertz Dipole	325
4.2	Filled Apertures	326
4.2.1	Angular Resolution and Efficiencies	326
4.2.2	Foci, Blockage, and Surface Accuracy	328
5	<i>Interferometers and Aperture Synthesis</i>	330
5.1	Aperture Synthesis	332
5.2	Interferometer Sensitivity	332
6	<i>Design Criteria</i>	332
6.1	Frequency Range	333
6.2	Sensitivity and Survey Speed	333
6.3	Angular Resolution	334
6.4	Field of View (FoV)	334

7	<i>The Antenna Arrays</i>	334
7.1	Fourier Synthesis Imaging	334
7.2	Crosses, Ts, and Other 2D Aperture Arrays	335
7.3	Phased Array Beamforming	336
7.4	Cylindrical Reflectors	336
7.5	Phased Array Feeds	337
7.6	Mosaicing	337
7.7	Rotation Measure Synthesis	338
7.8	Long Baseline Interferometry	338
8	<i>The Fundamental Differences Between Arrays and Dishes</i>	339
8.1	Filling Factor	339
8.2	Analog Beam Formation in the Focal Plane	339
8.3	Equivalence of Dishes and Arrays	339
8.4	Array Sensitivity	340
9	<i>Backends, Data Analysis, and Software</i>	341
10	<i>Types of Radio Frequency Interference (RFI) and Mitigation Strategies</i>	341
10.1	Radio Frequency Interference (RFI)	341
10.2	RFI Mitigation Methods	341
10.3	Adaptive Beam Nulling	342
11	<i>General Discussion</i>	342
11.1	Open Skies Policy in Radio Astronomy	342
11.2	Selecting the Best Telescope for Your Experiment	343
11.3	Analog Versus Digital	343
11.3.1	Fully Digital Receivers	344
11.4	General Purpose Versus Specialized Telescope Designs	344
12	<i>The World's Major Radio Telescopes</i>	345
12.1	Very Large Array (VLA, Now JVLA)	345
12.2	Australia Telescope Compact Array (ATCA)	346
12.3	VLBA (Very Long Baseline Array)	347
12.4	MERLIN (Multi-element Radio Linked Interferometer Network)	347
12.5	Parkes	347
12.6	Arecibo	348
12.7	Effelsberg 100-m Telescope	349
12.8	Green Bank Telescope (GBT)	350
12.9	Westerbork Synthesis Radio Telescope (WSRT)	350
12.10	Jodrell Bank	350
12.11	Giant Meterwave Radio Telescope (GMRT)	351
13	<i>Future Big Science Projects in Radio Astronomy</i>	351
13.1	The Karl G Jansky Very Large Array (JVLA, Previously the EVLA)	351
13.2	ALMA	351
13.3	LOFAR	351

13.4	Murchison Widefield Array (MWA)	352
13.5	Long Wavelength Array (LWA)	352
13.6	FAST	352
13.7	SKA and the SKA Precursors	352
13.7.1	MeerKAT	352
13.7.2	ASKAP	353
13.7.3	SKA	353
14	<i>The Future</i>	353
	<i>Appendix</i>	354
A.1.	Optical and Radio Analogs and Terminology	354
A.2.	The World's Largest Centimeter and Meter Radio Telescopes	354
	<i>References</i>	357

Abstract: “Radio Telescopes” starts with a brief historical introduction from Jansky’s 1931 discovery of radio emission from the Milky Way through the development of radio telescope dishes and arrays to aperture synthesis imaging. It includes sufficient basics of electromagnetic radiation to provide some understanding of the design and operation of radio telescopes. The criteria such as frequency range, sensitivity, survey speed, angular resolution, and field of view that determine the design of radio telescopes are introduced. Because it is so easy to manipulate the electromagnetic waves at radio frequencies, radio telescopes have evolved into many different forms, sometimes with “wire” structures tuned to specific wavelengths, which look very different from any kind of classical telescope. To assist astronomers more familiar with other wavelength domains, the [▶ appendix A.1.](#) includes a comparison of radio and optical terminology. Some of the different types of radio telescopes including the filled aperture dishes, electronically steered phased arrays, and aperture synthesis radio telescopes are discussed, and there is a section comparing the differences between dishes and arrays. Some of the more recent developments including hierarchical beam forming, phased array feeds, mosaicing, rotation measure synthesis, digital receivers, and long baseline interferometers are included. The problem of increasing radio frequency interference is discussed, and some possible mitigation strategies are outlined.

The open-sky policy adopted by most radio astronomy observatories makes it possible to select the best radio telescope for an experiment, and some guidelines are provided together with an [▶ appendix A.2.](#) listing all of the world’s large centimeter and meter radio wavelength telescopes. Finally, we include a short description of some of the great radio telescopes which have had the most scientific impact in the last decade and give some indications of future directions.

Keywords: Angular resolution, Aperture synthesis, Digital receivers, Field of view, History, Phased arrays, Radio telescopes, RFI, Sensitivity

1 Introduction

The intention is to give an overview of the history, design, and use of radio telescopes with emphasis on the general principals rather than detailed analysis. This should allow the astronomer to make informed decisions about the most suitable type of radio telescope to use to obtain observational information. Once this decision is made, most observatories will provide the tools needed to fine-tune sensitivity requirements and the other detailed specifications needed to make effective observations. These fundamentals are also essential when considering the design of future radio telescopes. The discussions of the basics is extended, and receivers, backends, and data processing are discussed in Volume 2, [▶ Chap. 6.](#)

Volume 1, [▶ Chap. 7,](#) covers submillimeter and millimeter radio telescopes, and we focus here on the wavelength range from about 30 GHz (1 cm) down to the ionospheric cutoff at about 15 MHz (20 m). This range of 2,000:1 in wavelength results in a great diversity in radio telescope designs.

Because it is so easy to manipulate the electromagnetic waves at radio frequencies, radio telescopes have evolved into many different forms, sometimes with “wire” structures tuned to specific wavelengths, which look very different from any kind of classical telescope. The most striking difference is between the dishes which are single monolithic collectors, either parabolic or parabolic sections, which concentrate the radiation at the focus where it is amplified in radio receivers, and the arrays which involve many separated receptors which sample the wave

front over large areas in the aperture plane and form an image by combining these signals electronically. In either case, the receiver amplifies the power collected by the telescope over a solid angle determined by the aperture of the telescope and the wavelength.

Appendix 1 summarizes some of the key concepts in optical and radio telescopes with the different nomenclature used at optical and radio wavelengths.

Appendix 2 is a compilation of the world's large cm radio telescopes. Only operating radio telescopes with a diameter greater than 25 m (or equivalent area) are included.

2 History

2.1 Early History

The definitive history of the development of radio astronomy from the beginnings in the 1930s up to 1953 can be found in Sullivan (2009). The new radio window on the universe was opened unexpectedly in 1932 by Jansky (1933a, b) from Bell Telephone Laboratory in the USA. Jansky detected radio emission from the Milky Way while investigating the source of noise on Trans Atlantic telephone routes. With no theoretical framework to understand these results and a vast communication gap between the astronomers and the engineers, this discovery was largely ignored until 1937 when Reber built the first parabolic dish larger than a few meters (see Reber 1958). Reber changed from the “bent wire” antennas to a 31-ft parabolic dish with amplifiers at the focus so he could work at higher frequencies (3.3 GHz) and more easily change frequency. At the time, the only concept for the source of the radio emission was thermal processes, so in this Rayleigh-Jeans region of the spectrum, it was expected that the emission detected by Jansky at 20.5 MHz (15 m) would be much more intense at cm wavelengths. This division between arrays of tuned elements at low frequencies and the more recognizable parabolic dishes at higher frequencies continues to this day.

During World War II the radio frequency technology developed rapidly due to the use of radar systems, so by 1946, the time was ripe to return to observations of the sky with far more sensitive radio equipment. Development of the equipment needed for receiving radio waves took off in many countries (UK, France, Australia, Japan, Russia, Canada, USA, see Sullivan (2009), but early developments were dominated by the groups led by Ryle at the University of Cambridge, UK, and by Pawsey at the CSIR (later CSIRO) in Sydney, Australia.

2.2 Evolution of Radio Telescope Sensitivity

2.2.1 Exponential Growth in Science

Harwit (1981) showed that the most important discoveries in astronomy result from technical innovation. The discoveries peak soon after new technology appears, and typically within 5 years of the technical capability. Instruments used for discoveries are often built by the observer. It had already been well established that most scientific advances follow technical innovation in other areas of science. de Solla Price (1963) applied quantitative measurement to the progress of science (scientometrics) and reached the conclusion that most scientific advances follow laboratory experiments. His analysis also showed that the normal mode of growth of science is exponential. A rather simplified conclusion to draw from this is that any

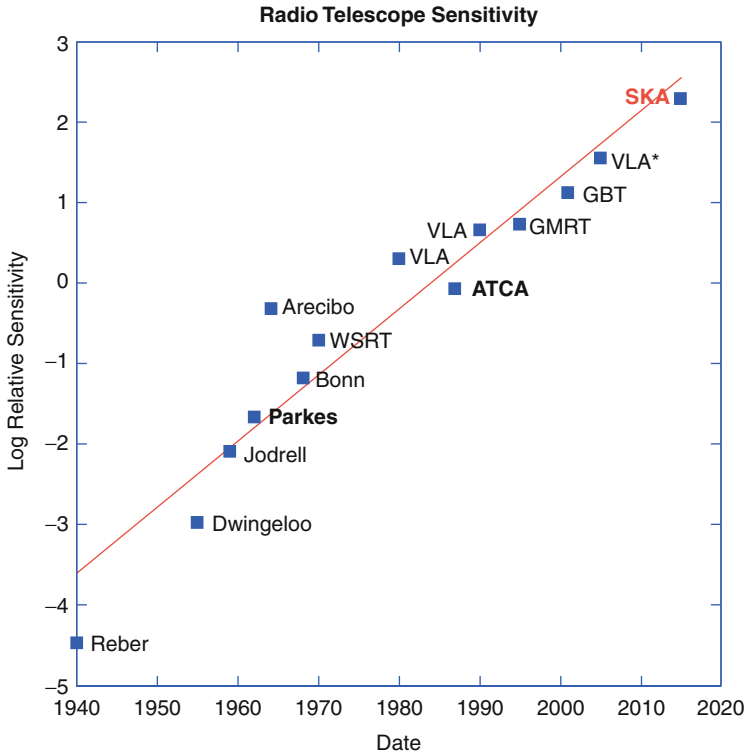


Fig. 8-1

Radio telescope sensitivity vs. time. Points are the relative continuum point source sensitivity when the telescopes were built, or after major upgrades. VLA* is the EVLA upgrade, now named the Jansky VLA. SKA is the proposed sensitivity for a telescope which has not yet been built (see Sect. 13)

field which has not maintained an exponential growth has now died out, so current active research areas are all still in an exponential growth phase. Furthermore, to maintain the exponential, the continual introduction of new technology is required since just refining existing technology plateaus out.

2.2.2 Livingston Curve

A famous example which illustrates this very well is the rate of increase of operating energy in particle accelerators by Livingston and Blewett (1962). Starting in 1930, each particle accelerator technology provided exponential growth up to a ceiling when a new technology was introduced. The envelope of the set of curves is itself an exponential with an increase in energy of 10^{10} in 60 years. This has been updated by Riesselmann (2009) to include the Large Hadron Collider. This example of exponential growth, originally presented by Fermi in 1954, has become known as the “Livingston Curve.”

To this we can add the now famous “Moore’s law” for computing devices (more precisely for transistors on a chip). Moore (1965) noted that the transistor density of semiconductor chips

doubled roughly every 1–2 years. This was later refined to doubling every 18 months, and this exponential growth has been maintained for the last 40 years (Mollick 2006).

➤ *Figure 8-1* plots the point source continuum sensitivity of telescopes used for radio astronomy since the first discovery of extraterrestrial radio emission in 1940. It has been exponential with an increase in sensitivity of 10^5 since 1940, doubling every 3 years. Also in this case, we can see particular radio telescope technologies reaching ceilings and new technologies being introduced, e.g., the transition from huge single dishes to arrays of smaller dishes in the 1980s.

2.3 The Development of the Aperture Synthesis Radio Telescope

Because of the long radio wavelengths, it was realized that interferometers with large spacings between the elements would be required to obtain high-enough angular resolution to determine the origin of the radio waves. Two of the main pioneering groups were at the University of Cambridge in the UK, led by Ryle, and at the CSIRO (then CSIR) Division of Radiophysics in Sydney, Australia, led by Pawsey. Both groups used the WWII radar technology to build astronomical instruments.

2.3.1 Australian Group

In Australia, the main focus was on solar imaging (see ➤ Chap. 71). The sun is a strong source but has a complex and time variable structure requiring good instantaneous measurements of the Fourier components. For this reason, the Australian arrays followed an evolutionary path with large numbers of relatively small elements.

In 1951, Christiansen built the Potts Hill grating array with thirty-two 6-ft diameter dishes near Sydney, Australia. By 1955, the first earth rotation synthesis image was obtained by Christiansen and Warburton (1955). Wild (1967) built a 3-km diameter circle of ninety-six 3-m dishes which made moving images of the radio sun. It operated for 17 years from 1967 and resolved many of the questions about the nature of solar bursts.

2.3.2 Cambridge Group

After early experiments observing the sun, Ryle's group in Cambridge moved their focus to the observation of "radio stars." These sources were static but much weaker than the sun, so the arrays evolved along a different path. They used movable antennas and earth rotation to build up the Fourier components over time, and they needed much larger elements to achieve the sensitivity required for the fainter sources. This evolution culminated in the construction of the One-Mile Telescope in 1963 (Ryle and Hewish 1960).

2.3.3 The Beginnings of Aperture Synthesis

The first published suggestion that it would be possible to synthesize an image of the radio sky by measuring a range of Fourier components was made by McCready et al. (1947). However, this technique was impractical with the cliff interferometers they were using and was not suitable for imaging solar bursts which were strongly variable in both time and frequency.

The first observations using a range of Fourier components measured with an interferometer with movable elements were made at the Cavendish Laboratory by Stanier (1950), Machin (1951), and O'Brien (1953).

2.3.4 Earth Rotation Synthesis

In June 1961, radio astronomers at the Cavendish Laboratory in Cambridge, UK, used 4C aerials operating at 178 MHz to make a radio source survey of the North Pole region using the earth's rotation to fully sample the aperture plane (Ryle and Neville 1962). Computations and graphical display used EDSACII which was the first use of a digital computer for radio astronomy imaging. The X-ray crystallographers in the Cavendish laboratory had developed the necessary Fourier transform programs. This was 7 years after Christiansen and Warburton (1955) first demonstrated the Earth's rotation synthesis with an observation of the quiet sun using an array of small dishes in Australia. However, the Australian group took many months to calculate by hand the Fourier transforms for one image, and the method was considered impractical at the time.

In 1962, the Cambridge group went on to build the One-Mile Telescope (Ryle and Hewish 1960) and the 5-km telescope in 1971. Ryle was awarded the Nobel Prize in 1974 "For his observations and inventions, in particular for the aperture synthesis technique."

The further development of Fourier imaging, deconvolution, and self-calibration will be described in Volume 2, [Chap. 7](#).

3 Radio Astronomy Fundamentals

A summary of the basics of electromagnetic radiation that are useful for the discussions in this chapter is provided in the following section. More details and physical background are available in a number of textbooks, for example, Kraus (1986), Burke and Graham-Smith (1996), and Wilson et al. (2008).

3.1 Radiative Transfer and Black Body Radiation

The total flux of a source is obtained by integrating intensity (in $\text{W m}^{-2} \text{Hz}^{-1} \text{sr}^{-1}$) over the total solid angle Ω_s subtended by the source

$$S_\nu = \int_{\Omega_s} I_\nu(\theta, \varphi) \cos \theta \, d\Omega. \quad (8.1)$$

The flux density of an astronomical source is given in units of a jansky (Jy). The jansky was adopted by the IAU in 1973 as the unit of spectral flux density. $1 \text{ Jy} = 10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1}$. The strongest of the (nonsolar) continuum radio sources are a few hundred jansky, and the current sensitivity limits for modern radio telescopes are now at the sub mJy level. Future telescopes such as the SKA will reach μJy sensitivity.

The spectral distribution of the radiation of a black body in thermodynamic equilibrium is given by the Planck law:

$$B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1} . \quad (8.2)$$

If $h\nu \ll kT$, the *Rayleigh-Jeans Law* is obtained:

$$B_{\text{RJ}}(\nu, T) = \frac{2\nu^2}{c^2} kT . \quad (8.3)$$

In the Rayleigh-Jeans relation, the brightness and the thermodynamic temperatures of black body emitters are strictly proportional (8.3). This feature is useful, so the normal expression of brightness of an extended source is *brightness temperature* T_B :

$$T_B = \frac{c^2}{2k} \frac{1}{\nu^2} I_\nu = \frac{\lambda^2}{2k} I_\nu . \quad (8.4)$$

If I_ν is emitted by a black body and $h\nu \ll kT$, then (8.4) gives the thermodynamic temperature of the source, a value that is independent of ν . If other processes are responsible for the emission of the radiation (e.g., synchrotron, free-free, or broadband dust emission), T_B will depend on the frequency; however, (8.4) is still used.

3.2 The Nyquist Theorem and Noise Temperature

This theorem relates the thermodynamic quantity temperature to the electrical quantities voltage and power. This is essential for the analysis of noise in receiver systems. The average power per unit bandwidth, P_ν (also referred to as power spectral density, PSD), produced by a resistor R is

$$P_\nu = \langle i\nu \rangle = \frac{\langle v^2 \rangle}{2R} = \frac{1}{4R} \langle v_N^2 \rangle , \quad (8.5)$$

where $v(t)$ is the voltage that is produced by i across R , and $\langle \dots \rangle$ indicates a time average. The first factor $\frac{1}{2}$ arises from the condition for the transfer of maximum power from R over a broad range of frequencies. The second factor $\frac{1}{2}$ arises from the time average of v^2 . Then

$$\langle v_N^2 \rangle = 4R k T . \quad (8.6)$$

When inserted into (8.5), the result is

$$P_\nu = k T . \quad (8.7)$$

8 Equation 8.7 can also be obtained by a reformulation of the Planck law for one dimension in the Rayleigh-Jeans limit. Thus, the available noise power of a resistor is proportional to its temperature, the *noise temperature* T_N , independent of the value of R and of frequency.

Not all circuit elements can be characterized by thermal noise. For example, a microwave oscillator can deliver $1 \mu\text{W}$, the equivalent of more than 10^{16} K, although the physical temperature is ~ 300 K. This is an example of a very *nonthermal* process, so temperature is not a useful concept in this case.

3.3 Overview of Intensity, Flux Density, and Main Beam Brightness Temperature

Temperatures in radio astronomy have given rise to some confusion. A short summary is given here. Power is measured by an instrument consisting of an antenna and a receiver. The power input can be calibrated and expressed as flux density or intensity. For very extended sources, intensity (see 8.4) can be expressed as a temperature, the *main beam brightness temperature*, T_{MB} . For discrete sources, the combination of (8.1) with (8.4) gives

$$S_{\nu} = \frac{2k}{\lambda^2} T_{\text{B}} \quad (8.8)$$

For a source with a Gaussian spatial distribution, this relation is

$$\left[\frac{S_{\nu}}{\text{Jy}} \right] = 0.0736 T_{\text{B}} \left[\frac{\theta}{\text{arcsec}} \right]^2 \left[\frac{\lambda}{\text{mm}} \right]^{-2} \quad (8.9)$$

if the flux density S_{ν} and the actual (or “true”) source size are known, then the *true brightness temperature*, T_{B} , of the source can be determined. For local thermodynamic equilibrium (LTE), T_{B} represents the physical temperature of the source. If the *apparent* source size, that is, the source angular size as measured with an antenna is known, (8.9) allows a calculation of T_{MB} . For discrete sources, T_{MB} depends on the angular resolution. If the antenna beam size has a Gaussian shape θ_{b} , the relation of actual θ_{s} and apparent size θ_{o} is:

$$\theta_{\text{o}}^2 = \theta_{\text{s}}^2 + \theta_{\text{b}}^2 \quad (8.10)$$

then from (8.8), the relation of T_{MB} and T_{B} is

$$T_{\text{MB}} (\theta_{\text{s}}^2 + \theta_{\text{b}}^2) = T_{\text{B}} \theta_{\text{s}}^2 \quad (8.11)$$

Finally, the PSD entering the receiver (8.7) is antenna temperature, T_{A} ; this is relevant for estimating signal to noise ratios (see (8.21) and (8.24)).

3.4 Polarization

Hertz dipoles are sensitive to a single linear polarization. By rotating the dipole over an angle perpendicular to the direction of the radiation, it is possible to determine the amount and angle of linearly polarized radiation. Helical antennas or arrangements of two dipoles are sensitive to circular polarization. Generally, polarized radiation is a combination of linear and circular, and is usually less than 100% polarized; so four parameters must be specified. It is usual to characterize polarization by the four Stokes parameters, which are the sum or difference of measured quantities. The total intensity of a wave is given by the parameter I . The amount and angle of linear polarization are given by the parameters Q and U , while the amount and sense of circular polarization are given by the parameter V . The definition of the sense of circular polarization in radio astronomy is the same as in electrical engineering but *opposite* to that used in the optical range; see Born and Wolf (1965) for a complete analysis of polarization, using the *optical* definition of circular polarization. Poincaré introduced a representation that permits an easy visualization of all the different states of polarization of a vector wave. See Radhakrishnan (1990), Thompson et al. (2001), or Wilson et al. (2008) for more details.

3.5 Sensitivity

The noise contributions from source, atmosphere, ground, telescope surface, and receiver are always additive:

$$T_{\text{sys}} = \sum T_i \quad (8.12)$$

From Gaussian statistics, the root mean square, RMS, noise is given by the mean value divided by the square root of the number of samples. From the estimate that the number of samples is given by the product of receiver bandwidth multiplied by the integration time, the result is

$$\Delta T_{\text{RMS}} = \frac{T_{\text{sys}}}{\sqrt{\Delta\nu\tau}} = \frac{(T_A + T_{\text{RX}})}{\sqrt{\Delta\nu\tau}} \quad (8.13)$$

Where T_A represents the power entering the receiver from the antenna and includes the source, the atmosphere, and the ground and T_{RX} represents the noise power added by the receiver. At very long (m) and short (mm) wavelengths, T_A will dominate T_{RX} . A more detailed derivation is to be found in Wilson et al. (2008).

4 Antennas

The antenna serves to focus power into the feed, a device that efficiently transfers power in the electromagnetic wave to the receiver. According to the principle of *reciprocity*, the properties of antennas such as beam sizes, efficiencies, etc. are the same whether these are used for receiving or transmitting. Reciprocity holds in astronomy, so it is usual to interchangeably use expressions that involve either transmission or reception when discussing antenna properties. For example, the terms *beam* and *feed* come from early radar usage. All of the following applies to the far-field radiation.

4.1 The Hertz Dipole

The total power radiated from a dipole carrying an oscillating current I at a wavelength λ is

$$P = \frac{2c}{3} \left(\frac{I\Delta l}{2\lambda} \right)^2 \quad (8.14)$$

For the dipole, the radiation is linearly polarized with the electric field along the direction of the dipole. The radiation pattern has a doughnut shape, with the cylindrically symmetric maximum perpendicular to the axis of the dipole. Along the direction of the dipole, the radiation field is zero. To improve directivity, reflecting screens have been placed behind a dipole, and in addition, collections of dipoles, driven in phase, are used. Dipole radiators have the best efficiency when the size of the dipole is $1/2\lambda$.

4.2 Filled Apertures

This section provides a simplified description of antenna properties. For more details, see Baars (2007). At cm and shorter wavelengths, flared waveguides (“feed horns”) or dipoles are used to convey power focused by the antenna (i.e., electromagnetic waves in free space) to the receiver (voltage). At the longest wavelengths, dipoles are used as the antennas. Details are to be found in Love (1976) and Goldsmith (1988, 1994).

4.2.1 Angular Resolution and Efficiencies

From diffraction theory (see Jenkins and White 2001), the angular resolution of a reflector of diameter D at a wavelength λ is

$$\theta = k \frac{\lambda}{D} \quad . \quad (8.15)$$

where k is of order unity. This universal result gives a value for θ (here in radians when D and λ have the same units). Diffraction theory also predicts the unavoidable presence of sidelobes, that is, secondary maxima. The sidelobes can be reduced by *tapering* the antenna illumination. Tapering lowers the response to very compact sources and increases the value of θ , that is, widens the beam due to the effective decrease in D .

The *normalized power pattern* is:

$$P_n(\vartheta, \varphi) = \frac{1}{P_{\max}} P(\vartheta, \varphi) \quad . \quad (8.16)$$

The *beam solid angle* Ω_A of an antenna is given by

$$\Omega_A = \int_{4\pi} \int P_n(\vartheta, \varphi) \, d\Omega = \int_0^{2\pi} \int_0^{\pi} P_n(\vartheta, \varphi) \sin \vartheta \, d\vartheta \, d\varphi \quad (8.17)$$

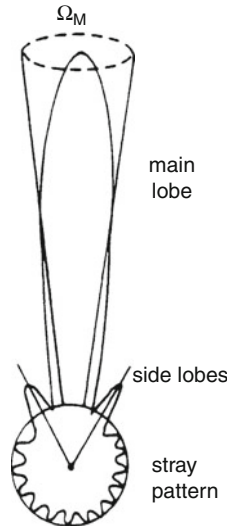
that is measured in steradians (sr). The integration is extended over all angles; so Ω_A is the solid angle of an ideal antenna having $P_n = 1$ for Ω_A and $P_n = 0$ everywhere else. For most antennas, the (normalized) power pattern has much larger values for a limited range of both ϑ and φ than for the remainder; the range where Ω_A is large is the main beam of the antenna; the remainder are the sidelobes or backlobes (► Fig. 8-2).

In analogy to (► 8.17), the *main beam solid angle* Ω_{MB} is defined as

$$\Omega_{MB} = \int \int_{\substack{\text{main} \\ \text{lobe}}} P_n(\vartheta, \varphi) \, d\Omega \quad . \quad (8.18)$$

The quality of a single antenna depends on how well the power pattern is concentrated in the main beam. The definition of *main beam efficiency* or *beam efficiency*, η_B , is

$$\eta_B = \frac{\Omega_{MB}}{\Omega_A} \quad . \quad (8.19)$$



■ Fig. 8-2

A polar power pattern showing the main beam, and near- and farside lobes. The weaker farside lobes have been combined to form the stray pattern

η_B which is the fraction of the power is concentrated in the main beam. The main beam efficiency can be modified (within limits) for parabolic antennas by changing the illumination of the main reflector. An under-illuminated antenna has a wider main beam but lower sidelobes. The angular extent of the main beam is usually described by the *full width to half power width* (FWHP), the angle between points of the main beam where the normalized power pattern falls to $1/2$ of the maximum. The beamwidth θ is given by (8.15).

If a plane wave with the power density $|\langle \vec{S} \rangle|$ in W m^{-2} is intercepted by an antenna, a certain amount of power is extracted from this wave. This power is P_e , and the fraction is

$$A_e = P_e / |\langle \vec{S} \rangle|, \quad (8.20)$$

the *effective aperture* of the antenna. A_e has the dimension of m^2 . Compared to the *geometric aperture* A_g , an aperture efficiency η_A can be defined by

$$\boxed{A_e = \eta_A A_g} \quad . \quad (8.21)$$

If an antenna with a normalized power pattern $P_n(\vartheta, \varphi)$ is used to receive radiation from a brightness distribution $B_\nu(\vartheta, \varphi)$ in the sky, at the output terminals of the antenna, the power per unit bandwidth (PSD), in W Hz^{-1} , P_ν is

$$P_\nu = \frac{1}{2} A_e \int \int B_\nu(\vartheta, \varphi) P_n(\vartheta, \varphi) d\Omega . \quad (8.22)$$

By definition, this operates in the Rayleigh-Jeans limit, so the equivalent distribution of brightness temperature can be replaced by an equivalent *antenna temperature* T_A (8.7):

$$P_\nu = k T_A . \quad (8.23)$$

This definition of *antenna temperature* relates the output of the antenna to the power from a matched resistor. When these two power levels are equal, then the antenna temperature is given by the temperature of the resistor. The effective aperture A_e can be replaced by the the beam solid angle $\Omega_A \cdot \lambda^2$. Then (8.22) becomes

$$T_A(\vartheta_0, \varphi_0) = \frac{\int T_B(\vartheta, \varphi) P_n(\vartheta - \vartheta_0, \varphi - \varphi_0) \sin \vartheta \, d\vartheta \, d\varphi}{\int P_n(\vartheta, \varphi) \, d\Omega} \quad (8.24)$$

From (8.24), $T_A < T_B$ in all cases. The numerator is the *convolution* of the brightness temperature with the beam pattern of the telescope (Fourier methods are of great value in this analysis; see Bracewell 1986). The brightness temperature $T_b(\vartheta, \varphi)$ corresponds to the thermodynamic temperature of the radiating material *only* for thermal radiation in the Rayleigh-Jeans limit from an optically thick source; in all other cases, T_B is a convenient quantity that represents source intensity at a given frequency.

For a source small compared to the beam, (8.22) and (8.23) give

$$P_v = \frac{1}{2} A_e S_v = k T_A \quad (8.25)$$

T_A is the antenna temperature at the receiver.

$$T_A = \Gamma S_v \quad (8.26)$$

where Γ is the *sensitivity* of the telescope measured in K Jy^{-1} . Introducing the aperture efficiency η_A according to (8.21), we find

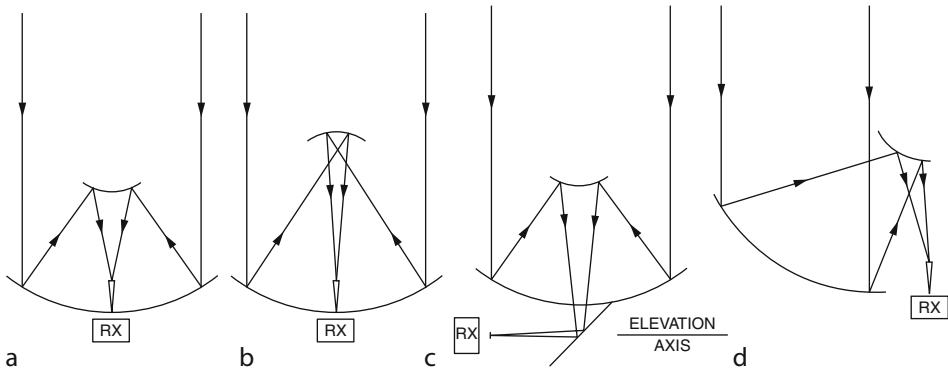
$$\Gamma = \eta_A \frac{\pi D^2}{8k} \quad (8.27)$$

Thus, Γ or η_A can be measured with the help of a calibrating source provided that the diameter D and the noise power scale in the receiving system are known. When (8.26) is solved for S_v , the result is:

$$S_v = 3,520 \frac{T_A(\text{K})}{\eta_A(\text{D/m})^2} \quad (8.28)$$

4.2.2 Foci, Blockage, and Surface Accuracy

If the size of a radio telescope is more than a few hundred wavelengths, designs are similar to those of optical telescopes. Cassegrain, Gregorian, and Nasmyth systems have been used. See Fig. 8-3 for a sketch of these focal systems. In a Cassegrain system, a convex hyperbolic reflector is introduced into the converging beam immediately in front of the prime focus. This reflector transfers the converging rays to a secondary focus which, in most practical systems, is situated close to the apex of the main dish. A Gregorian system makes use of a concave reflector with an elliptical profile. This must be positioned behind the prime focus in the diverging beam. In the Nasmyth system, this secondary focus is situated in the elevation axis of the telescope by introducing another, usually flat, mirror. The advantage of a Nasmyth system is that the receiver



■ Fig. 8-3

The geometry of parabolic apertures: (a) Cassegrain, (b) Gregorian, (c) Nasmyth, and (d) offset Cassegrain systems (From Wilson et al. 2008)

front ends remain horizontal while the telescope is pointed toward different elevations. This is an advantage for receivers cooled with liquid helium, which may become unstable when tipped. Cassegrain and Nasmyth foci are commonly used in the mm/sub-mm wavelength ranges.

In a secondary reflector system, feed illumination beyond the edge receives radiation from the sky, which has a temperature of only a few K. For low-noise systems, this results in only a small overall system noise temperature. This is significantly less than for prime focus systems. This is quantified in the so-called G/T value, that is, the ratio of antenna gain to system noise. Any telescope design must aim to minimize the excess noise at the receiver input while maximizing gain. For a specific antenna, this maximization involves the design of feeds and the choice of foci.

The secondary reflector and its supports block the central parts in the main dish from reflecting the incoming radiation, causing some significant differences between the actual beam pattern and that of an unobstructed antenna. Modern designs seek to minimize blockage due to the support legs and subreflector.

Feed leg blockage will cause deviations from circular symmetry. For altitude-azimuth telescopes, these sidelobes will change position on the sky with hour angle (see Reich et al. 1978). This may be a serious defect, since these effects will be significant for maps of low-intensity regions near an intense source. The sidelobe response may depend on the polarization of the incoming radiation. Equatorially mounted telescopes avoid this problem but at high cost for a large telescope. A new option now being explored is the three axis mount so an alt-az telescope can rotate about a third axis parallel to the beam axis keeping the sidelobe patterns fixed on the sky.

The gain of a filled aperture antenna with small-scale surface irregularities ϵ cannot increase indefinitely with increasing frequency but reaches a maximum at $\lambda_m = 4\pi\epsilon$, and this gain is a factor of 2.7 below that of an error-free antenna of identical dimensions. The usual rule-of-thumb is that the irregularities should be 1/16 of the shortest wavelength used.

5 Interferometers and Aperture Synthesis

From diffraction theory, the angular resolution is given by (8.15). However, as shown by Michelson (see Jenkins and White 2001), a much higher resolving power can be obtained by coherently combining the output of two reflectors of diameter $d \ll B$ separated by a distance B yielding $\theta \approx \lambda/B$. In the Rayleigh-Jeans regime, $h\nu \ll kT$, the outputs can be amplified without seriously degrading the signal-to-noise ratio. This amplified signal can be divided and used to produce a large number of cross-correlations (e.g., see Radhakrishnan 1999).

The simplest case is a two-element system in which electromagnetic waves are received by two antennas. These induce the voltage V_1 at A_1 :

$$V_1 \propto E e^{i\omega t}, \quad (8.29)$$

while at A_2 :

$$V_2 \propto E e^{i\omega(t-\tau)}, \quad (8.30)$$

where E is the amplitude of the incoming electromagnetic plane wave, τ is the geometric delay caused by the relative orientation of the interferometer baseline \vec{B} and the direction of the wave propagation. For simplicity, receiver noise and instrumental phase were neglected in (8.29) and (8.30). These two outputs will be correlated. Today, all radio interferometers use direct correlation followed by an integrator.

The output is proportional to

$$R(\tau) \propto \frac{E^2}{T} \int_0^T e^{i\omega t} e^{-i\omega(t-\tau)} dt.$$

If T is a time much longer than the time of a single full oscillation, that is, $T \gg 2\pi/\omega$, then the average over time T will not differ much from the average over a single full period, resulting in

$$R(\tau) \propto E^2 e^{i\omega\tau}. \quad (8.31)$$

The output of the correlator + integrator varies periodically with τ , the delay. Since \vec{s} is slowly changing due to the rotation of the earth, τ will vary, producing *interference fringes* as a function of time.

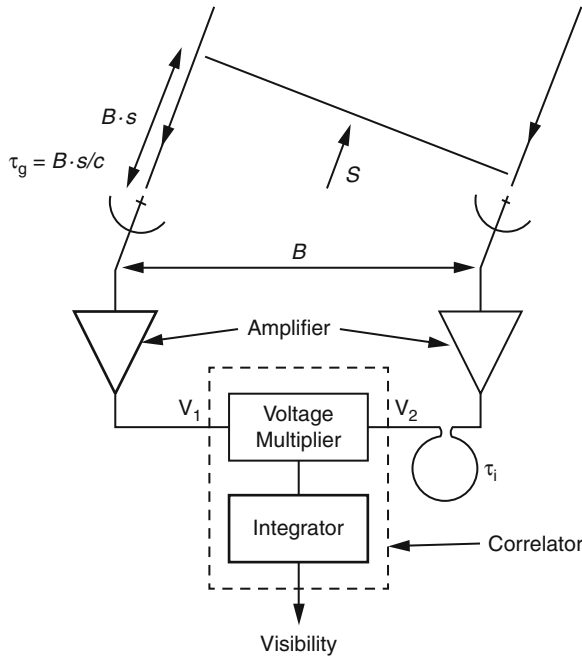
The basic components of a two-element system are shown in Fig. 8-4. If the radio brightness distribution is given by $I_\nu(\vec{s})$, the power received per bandwidth $d\nu$ from the source element $d\Omega$ is $A(\vec{s})I_\nu(\vec{s})d\Omega d\nu$, where $A(\vec{s})$ is the effective collecting area in the direction \vec{s} ; the same $A(\vec{s})$ is assumed for each of the antennas. The amplifiers are assumed to have constant gain and phase factors (neglected here for simplicity).

The output of the correlator for radiation from the direction \vec{s} (Fig. 8-4) is

$$r_{12} = A(\vec{s}) I_\nu(\vec{s}) e^{i\omega\tau} d\Omega d\nu \quad (8.32)$$

where τ is the difference between the geometrical and instrumental delays τ_g and τ_i . If \vec{B} is the baseline vector between the two antennas

$$\tau = \tau_g - \tau_i = \frac{1}{c} \vec{B} \cdot \vec{s} - \tau_i, \quad (8.33)$$



■ Fig. 8-4

A schematic diagram of a two-element correlation interferometer. The antenna output voltages are V_1 and V_2 ; the instrumental delay is τ_i and the geometric delay is τ_g . \vec{s} is the direction to the source. Perpendicular to \vec{s} is the projection of the baseline \vec{B} . The signal is digitized after conversion to an intermediate frequency. Time delays are introduced using digital shift registers (From Wilson et al. 2008)

the total response is obtained by integrating over the source S

$$R(\vec{B}) = \int_{\Omega} \int A(\vec{s}) I_v(\vec{s}) e^{2\pi i \nu (\frac{1}{c} \vec{B} \cdot \vec{s} - \tau_i)} d\Omega d\nu \quad (8.34)$$

The function $R(\vec{B})$, the *Visibility Function* is closely related to the mutual coherence function (see Born and Wolf 1965; Thompson et al. 2001; Wilson et al. 2008) of the source. For parabolic antennas, it is usually assumed that $A(\vec{s}) = 0$ outside the main beam area so that (8.34) is integrated only over this region. A one-dimensional version of (8.34), for a baseline B , frequency $\nu = \nu_0$ and instrumental time delay $\tau_i = 0$, is

$$R(B) = \int A(\theta) I_v(\theta) e^{2\pi i \nu_0 (\frac{1}{c} B \theta)} d\theta, \quad (8.35)$$

With $\theta = x$ and $B_x / \lambda = u$, this is

$$R(B) = \int A(\theta) I_v(\theta) e^{2\pi i u x} d\theta. \quad (8.36)$$

This form of (8.34) illustrates more clearly the Fourier transform relation of u and x .

5.1 Aperture Synthesis

To produce an image, the integral equation (8.34) must be inverted. A number of approximations may have to be applied to produce high-quality images. In addition, the data are affected by noise.

For imaging over a limited region of the sky, rectangular coordinates are adequate, so relation (8.34) can be rewritten with coordinates (x, y) in the image plane and coordinates (u, v) in the Fourier plane. The coordinate w , corresponding to the difference in height, is set to zero. Then the relevant relation is

$$I'(x, y) = A(x, y) I(x, y) = \int_{-\infty}^{\infty} V(u, v, 0) e^{-2\pi i (ux+vy)} du dv \quad (8.37)$$

where $I'(x, y)$ is the intensity $I(x, y)$ modified by the primary beam shape $A(x, y)$. It is easy to correct $I'(x, y)$ by dividing by $A(x, y)$.

5.2 Interferometer Sensitivity

The random noise limit to an interferometer system can be calculated following the method used for a single telescope (8.13). The use of (8.25) provides a conversion from ΔT_{RMS} to ΔS_v . For an array of n identical antennas, there are $N = n(n-1)/2$ simultaneous pairwise correlations, so the RMS variation in flux density is

$$\Delta S_v = \frac{2 M k T_{\text{sys}}}{A_e \sqrt{2 N t \Delta \nu}}, \quad (8.38)$$

with $M \geq 1$, A_e the effective area of each antenna and T_{sys} given by (8.12). This relation can be recast in the form of brightness temperature fluctuations using the Rayleigh-Jeans relation; then the RMS noise in brightness temperature units is

$$\Delta T_B = \frac{2 M k \lambda^2 T_{\text{sys}}}{A_e \Omega_b \sqrt{2 N t \Delta \nu}}. \quad (8.39)$$

For a Gaussian beam, $\Omega_{\text{mb}} = 1.133 \theta^2$, so the RMS temperature fluctuations can be related to observed properties of a synthesis image.

6 Design Criteria

Parabolic dishes are the design choice if a large collecting area is required with frequency agility. All high-frequency radio telescopes are based on the parabolic dish concept, either single dishes (8.4.2) or arrays of such dishes (8.7).

The advantage of the parabolic dish is the flexibility resulting from having a single receiver at the focal point. Frequency agility can then be obtained by changing the receiver system used. Such receiver changes can be made manually or with automatic systems involving rotating turrets, translators, or tilting subreflectors to move the different receivers into the focus.

However, recent developments are now changing this many decades old paradigm: Narrow-band highly optimized single pixel feeds which had to be moved in and out of the focal position

to change observing frequency are being replaced by wideband feeds and receivers which cover all the required frequencies simultaneously. The single pixel feeds which observed a single point on the sky are being replaced by multibeam receivers or focal plane arrays which can use more of the information in the focal plane to image a larger area of sky (🔗 Sect. 7.5) or to observe more than one direction simultaneously as in the VERA astrometric VLBI array (Kawaguchi et al. 2000). The huge advances in digital processing have changed the balance between analog and digital, moving the transition from the aperture arrays of dipoles to the parabolic dishes to higher frequencies (🔗 Sect. 11.3).

6.1 Frequency Range

The combination of the antenna and feed which couples the electromagnetic radiation field to the amplifying receiver makes it impossible to design a single detector system that works over the entire $>1,000:1$ radio astronomy wavelength range. High-efficiency systems are generally constrained to an octave bandwidth, but extending this range is an active current research topic in feed design, for example, Kildal et al. (2009), Akgiray et al. (2011). Bandwidths of 10:1 are now being achieved. The advent of higher-speed digital signal processing (🔗 Sect. 11.3.1) has also made it possible to increase the backend bandwidth to many GHz so a wide range of frequency can be observed simultaneously.

6.2 Sensitivity and Survey Speed

From (🔗 8.13) and (🔗 8.25) for one polarization channel, we have

$$\Delta S_v = \frac{2k(T_A + T_{rx})}{A_e \sqrt{\tau \Delta \nu}}.$$

The maximum possible point source sensitivity only depends on effective aperture area (A_e), system temperature ($T_{sys} = T_A + T_{rx}$), observing time (τ), and bandwidth ($\Delta \nu$). 🔗 Section 4.2.2 discussed the optimization of T_{sys} and A_e by minimizing “G/T.” Further details of radio telescope receivers will be discussed in Volume 2 🔗 Chap. 6 but we note that optimized modern radio receivers in the cm bands are now close to the ultimate noise limits, and in the meter band are already well below the background noise, $T_A > T_{rx}$. Hence, the only way to further improve point source sensitivity is with more collecting area, more bandwidth, or more observation time. For spectral line observations, more bandwidth may increase the number of lines observed but cannot improve the sensitivity for any one line. It is the extreme sensitivity needed to observe the unique 21 cm H line at cosmological distances that has driven the vision of a square kilometer area array (SKA) discussed in 🔗 Sect. 13.7.3. Similarly, for some transient phenomena, it is not possible to increase the observing time so again increasing collecting area is the only way forward.

For surveys, the limiting sensitivity in a given time also depends on the field of view, FoV. A survey speed figure of merit (SSFoM) can be defined which is related to the number of sources detected in a given time.

$$\text{SSFoM} \propto \text{FoV} \cdot \Delta \nu \cdot (A_e / T_{sys})^2 \quad (8.40)$$

where FoV = total instantaneous solid angle, $\Delta\nu$ = bandwidth, A_e = total effective area of the array, and T_{sys} = system temperature, Cordes (2007). This opens up other design opportunities as it is possible to make trade-offs between collecting area and field of view (see [Sects. 6.4](#) and [7.5](#)).

6.3 Angular Resolution

The struggle to obtain higher angular resolution has been the main driver since the beginning of radio astronomy. The development of all aspects of high angular resolution in radio astronomy has been reviewed by Kellermann and Moran (2001). It is ironic to note that although the long radio wavelength results in poor angular resolution, the ability to independently sample the electromagnetic field on almost unlimited baselines also gives the highest angular resolution in astronomy, see [Sect. 7.8](#).

6.4 Field of View (FoV)

The instantaneous FoV of any radio telescope can be shown to be the product of the diffraction-limited primary beam ([8.15](#)) and the number of independent receiving elements. These receiving elements can be either in the focal or the aperture plane. The receivers might be incoherent bolometer arrays (preferred at high frequency) or coherent detectors including both multibeam receivers and phased array feeds as discussed in the next section.

7 The Antenna Arrays

The basic theory of the interferometer and an introduction to aperture synthesis is given in [Sect. 5](#). Here we explore a number of topics which may be of interest to a broader multi-wavelength audience. There are some excellent references for the basics and for more detailed treatment, for example, Thompson et al. (2001), and Christiansen and Hogbom (1985). Further aspects of this topic are included in [Chap. 7](#) on Radio and Optical Interferometry.

7.1 Fourier Synthesis Imaging

The theory of interferometry is straightforward: the coherence of the radiation field is measured over a large volume of space and an image formed by Fourier inversion. The practical problems of measuring the coherence from telescopes confined to the earth's surface are numerous but not too onerous. A very useful collection of practical lectures on modern aperture synthesis techniques can be found in the NRAO Summer School, Taylor et al. (1999).

The radiation can be conveniently collected using parabolic reflectors with typical diameters at centimeter wavelengths in the range 5–100 m, a diameter of 25 m being a popular compromise between cost of construction and sensitivity. Both the ability to point the antennas and the surface accuracy limit the highest practicable observing frequency.

Once the signals have been collected and amplified, they must be relayed to a central location where they are correlated with each other to form estimates of the coherence function ([Sect. 5](#)). Direct digital correlation is almost universally preferred for the flexibility and

lower levels of systematic error. Before the correlation is performed, the signals must be lined up to eliminate the continually changing geometric delay of one antenna relative to the other due to the earth's rotation. This process is analogous to tracking of a celestial object by continually tilting an optical telescope as the earth rotates. For digitized signals, the delaying is conveniently performed using large digital buffers of high-speed memory. This earth rotation complicates matters somewhat. First, a given pair of antennas, as seen from the object, rotates around in the aperture plane, smearing out fine structure in the coherence function unless the integration time is short enough. The integration timescale in which the visibility is smeared is simply related to the time it takes for earth rotation to move the most distant antenna by its diameter. For a 25 m antenna at tens of kilometers baseline, this is about 10 s. Second, the relative motion between the antennas introduces a differential Doppler shift in the radiation received, which must be canceled before correlation.

High-quality imaging required good sampling of the coherence across the aperture plane. The largest separation of the antennas fixes the highest resolution possible, while the distribution of the samples over the aperture plane determines the complexity of structure that can be imaged. While the goal is to obtain the densest possible coverage of the aperture plane, the finite number of observed correlations and the desire for high angular resolution usually results in a compromise with incomplete sampling and higher sidelobes in the synthesis image. A detailed description of the range of deconvolution algorithms which correct for the nonuniform or incomplete sampling of the aperture plane is included in [▶ Chap. 7](#) on Radio and Optical Interferometry and in Volume 2 [▶ Chap. 6](#) on Techniques of Radio Astronomy.

The coherence measurements can be made at any time over the interval while the structure has not changed. This is of course the requirement for earth rotation synthesis, and it also makes it possible to build up samples of coherence at different spacings by moving the array elements. This is essential to obtain dense coverage of the aperture plane and high resolution with a small number of elements. Many aperture synthesis arrays have some, or all, antennas moveable (VLA, WSRT, ATCA in [▶ Sect. 12](#)).

7.2 Crosses, Ts, and Other 2D Aperture Arrays

The list of large cm radio telescopes in Table A2 includes many crosses and T-shaped aperture arrays. Ryle (1952) first pointed out that an effective gain in resolving power can be obtained with an interferometer consisting of two dissimilar primary antennae, for example, one with a narrow beam in the east-west direction and the other with a narrow beam in the north-south direction. The next logical step was made by Mills and Little (1953), who proposed that if the two elements had a common electrical and physical center (virtually an interferometer with separations down to zero spacing), a pencil beam antenna would result. Such an antenna is known as a Mills Cross. To minimize cost, simple collinear feed systems with fixed lengths from each element to the receiver were used but this greatly restricts the band width.

This design produced a high-quality pencil beam and could be implemented with all analog components, requiring minimal digital signal processing and no Fourier transform calculations. By the 1990s, digital technology had made this approach unattractive, and some of the old crosses and Ts were modified to incorporate the more flexible digital technology (e.g., Large et al. 1994). New array designs such as VLA ([▶ Sect. 12.1](#)), GMRT ([▶ Sect. 12.11](#)), ASKAP ([▶ Sect. 13.7.2](#)), and MEERKAT ([▶ Sect. 13.7.1](#)) use nonuniform 2D antenna element distributions to measure the maximum number of Fourier components with the minimum number of elements (minimum redundancy arrays) instead of the cross or T geometry. However, the VLA

still uses linear structures, the arms of the Y, to simplify antenna movement. In a recent twist, the computer processing and I/O requirements have become so extreme that telescope designs are again being proposed with geometrical arrangements of elements which minimize the computational load, for example, Tegmark and Zaldarriaga (2010) propose a regular antenna grid for optimum use of the fast Fourier transform, and Bunton (2011) proposes redundant geometries to decrease computer load.

7.3 Phased Array Beamforming

The traditional phased arrays formed a single beam by electronically adjusting the phases of all the elements to correspond to a given pointing direction. This is still the preferred telescope design at low frequencies because the effective area of a single antenna element, such as a simple dipole is $\sim \lambda^2$ so is large at long wavelengths, and the cost per element is low. Hence, sparse-phased arrays are a very cost-effective way to obtain a large collecting area at very low frequency. However, if only a single-phased array beam is formed, the advantage of the wide FoV of each element of the array is lost and the survey speed is greatly reduced. Since we are in the Rayleigh-Jeans regime with many photons per state in the system, for example, see Radhakrishnan (1999), we can split the signal without any loss in S/N to form multiple beams pointing in different directions. It is then theoretically possible to form enough simultaneous beams to cover the whole FoV of each element. However, the number of beams needed scales as $n = A_{\text{tot}}/\lambda^2$ for an n element array of total geometrical area A_{tot} . The inputs to the electronic beamformer will also scale with n , so some of the electronics is scaling as $n^2 (A_{\text{tot}}/\lambda^2)^2$. For 1 km² array at 1 m, this is a factor of 10^{12} which is well beyond even Moore's law extrapolation and beyond viable power consumption limits. The solution to this dilemma is to introduce hierarchical beamforming which allows a smooth trade-off between the number of beams, and hence electronics cost, and the areas of sky imaged simultaneously. In an array with hierarchical beam formation such as LOFAR (☛ Sect. 13.3), groups of elements, referred to as tiles, are combined to form the *tile beam*, groups of tiles in one geographic location are then combined to form one or more *station beams* and the stations are finally combined using the normal cross-correlating techniques of aperture synthesis to form a *synthesized beam*. Hence, an observation with LOFAR will have a tile beam (roughly equivalent to the *primary beam* of an array of single dishes), a number of *station beams* which can be simultaneously pointing to different regions in the *tile beam*, and finally a Fourier transform image for the FoV of each *station beam* with the resolution of the *synthesized beam*.

Hierarchical beamforming in phased arrays is used to reduce the FoV and information data rate to a manageable level. In this process, it is possible to replace parts of the digital hierarchy with analog beamforming, and the cylinders and phased array feeds in dishes discussed in the next two sections are two examples.

7.4 Cylindrical Reflectors

Another method of reducing the FoV is to place a one-dimensional phased array at the focus of a parabolic cylindrical reflector. For a given collecting area, the reduction in the number of inputs to the beamformer is reduced by the factor $\lambda/2D$ where D is the width of the cylindrical reflector. Typically, this is a factor of 0.1–0.01 allowing the cylinders to operate at

higher frequencies than phased arrays for the same electronics cost. The advantage of the cylinder over a dish is the fact that the reflector is much cheaper, and this made the approach popular in the 1960s. However, the number of feed elements in one-dimensional phased arrays make cylindrical reflectors hard to upgrade and they were gradually displaced by the dishes.

7.5 Phased Array Feeds

There is an exact analogy between phased array beamforming in the aperture plane and the phased array feeds (PAFs) which can be placed in (or near) the focal plane of a single dish. The great advantage of putting the PAFs at the focus of a single dish is to reduce the number of active elements while retaining the sensitivity corresponding to the area of the dish. The next logical step is to place arrays of feeds in the focal plane of each dish in an interferometer array to image many fields at once. First suggested by Fisher and Bradley (2000) and now being implemented in Westerbork (APERTIF) based on an array of Vivaldi feeds (Oosterloo et al. 2010) and in ASKAP based on a self-complimentary checker board antenna using a connected patch array (Hay et al. 2008). Note that for the same FoVm the electronics' cost advantage of the PAF over the arrays is area PAF/area dish which is in the range 0.01–0.001 for current designs.

The PAF can be placed at any location along the wavefront as long as the complex distribution of amplitude of the electromagnetic wave is fully sampled. The PAF does not have to be in the focal plane, but as long as it is in a region where the wavefront has a small waist, the number of receptor elements needed is minimized.

PAFs are also referred to as focal plane arrays (FPA). The terminology PAF is preferred for arrays of elements which are combined to form phased array beams. The term FPA would refer to any array of receiving elements in the focal plane, including multibeam receivers which do not sample the focal plane continuously and bolometer detector arrays which produce separate total power beams for each element. The term “smart feeds” is also used for feeds which are electronically configurable.

7.6 Mosaicing

When a number of interferometric imaging observations are made with overlapping primary beams to cover a larger area of the sky, this is called a mosaic. Mosaicing has two benefits; obviously, the area of sky observed is now larger than the primary beam, but more subtly, additional Fourier components of the sky brightness distribution have been measured (Ekers and Rots 1979). To see how this is possible, first consider a single pointing made with a filled aperture. All the Fourier components, from the area of sky in its primary beam, are combined into a single value with weights determined by the illumination of the aperture (taper). If two dishes form an interferometer all the Fourier components corresponding to all possible pairs of elements between the two apertures are combined into a single complex visibility, that is, for two dishes of diameter D , separated by a baseline B , the resulting visibility is the average of baselines from $B - D$ to $B + D$. By combining two interferometer observations with different pointings with overlapping primary beams, all the Fourier components from $B - D$ to $B + D$ are recovered. This is particularly powerful when B is $\approx D$ because the recorded baselines from 0 to $2D$ are the “missing” short spacings that plague interferometric observations of large sources. A simple way to see how this works is to consider the two different pointings as two different phase gradients

across the aperture of the dishes. These known phase gradients make it possible to disentangle the combined Fourier components. Cornwell (1988) and Cornwell et al. (1993) demonstrate how the image deconvolution algorithms can be modified to incorporate the additional information measured in a mosaiced observation. Methods using both the linear combination of overlapping primary beams and the joint deconvolution of overlapping primary beams are now in routine use.

A new and exciting development is the combination of the use of phased array feeds in each element of the interferometer array. Now the overlapping mosaiced fields are all measured simultaneously, greatly increasing the instantaneous FoV but also reducing the effect of some errors caused by the changing atmosphere or single dish pointing, which accumulate in sequential observations of a mosaiced area.

Mosaicing becomes especially important at higher frequencies when the primary beam of the larger and more sensitive telescopes is relatively narrow. For a more detailed treatment of mosaicing, see Sect. 11.6 in Thompson et al. (2001).

7.7 Rotation Measure Synthesis

The birefringence of the magnetized plasma in interstellar and intergalactic space causes the observed linear polarization properties of radio sources to be strongly frequency-dependent. This Faraday rotation causes the position angle of the linear polarization vector to increase by an amount $RM \cdot \lambda^2$ where the magnitude of the rotation measure (RM) ranges between 10 rad m^{-2} for typical sources observed through the interstellar medium of our Galaxy and $5 \times 10^5 \text{ rad m}^{-2}$ for the compact radio source at the center of the Milky Way. Observations made over a large frequency range will average out all the linear polarization, unless the spectral resolution is fine enough that Faraday rotation does not cause appreciable rotation of the polarization vector across individual spectral channels (bandwidth smearing). However, the signal-to-noise ratio of the polarization measurement in each spectral channel may be too low to compute the polarization vector. The technique of rotation measure (RM) synthesis has been developed, for example, Brentjens and de Bruyn (2005), to simultaneously utilize the measurements across an entire wide frequency band and is the optimum method to extract polarization information from noisy data Macquart et al. (2012).

7.8 Long Baseline Interferometry

In 1967, a new technique of interferometry was developed in which the receiving elements were separated by such a large distance that it was necessary to operate them independently with no real-time communications link. This was accomplished by recording the undetected voltages from each site on magnetic tape timed using independent atomic clocks sufficiently accurate to maintain coherence. Later, this data is cross-correlated at a central processing station. The technique is called very long baseline interferometry (VLBI). The principles involved in VLBI are fundamentally the same as those involved in interferometers with connected elements (Thompson et al. 2001). We are now seeing a convergence of VLBI and connected arrays as the independent tape recorders are being replaced by wideband fiber optic communications links, and it is now also technically possible to maintain phase coherence with stabilized optical links.

8 The Fundamental Differences Between Arrays and Dishes

8.1 Filling Factor

A filled aperture telescope measures all Fourier components up to its maximum diameter and has brightness sensitivity which is independent of diameter. An array has a filling factor, η , which is less than 1. The brightness sensitivity of the array is decreased, and the power in its side-lobe increased by $1/\eta$. The array may still measure all Fourier components up to its maximum spacing and consequently have relatively low sidelobes even though η is less than 1 because the number of redundant spacings can be minimized. However, high angular resolution arrays will often have η much less than 1 and incomplete coverage of Fourier components with higher sidelobes and lower brightness sensitivity.

8.2 Analog Beam Formation in the Focal Plane

A conventional radio telescope at shorter wavelengths, for example, single parabolic dish with a conventional focal plane, will form a beam when the signals reflected from all the apertures are combined at a point in the focal plane before amplification and detection. The parabolic shape of the dish surface ensures that all signals from the pointing direction are combined in phase. In an array, the undetected receiver voltage outputs from each element are sampled and stored before they are later combined with appropriate delay and phase for each direction in the sky. This corresponds to the Fourier transform operation as described in [Sects. 5](#) and [7.1](#), and an aperture array is normally performed in a digital computer.

8.3 Equivalence of Dishes and Arrays

Consider a single parabolic dish of diameter, d , focusing radiation from a given direction in the sky at a point in the focal plane. The surface of the dish can be divided into n contiguous sub elements of area $A_i = A_{\text{tot}}/n$. The parabolic shape of the dish ensures that the path length to the focus will be the same for radio waves reflected from each element of the aperture. Now replace each aperture element by a small dish and receiver system which samples and amplifies the voltage averaged over the same area A_i . The voltages from all these sub elements, V_i , can then be added in the phase (either in real time or later in a computer) to produce the same signal that would be received at the focus of the single dish. It can then be detected $(\sum V_i)^2$ to obtain the power from the direction in which the dish is pointing. Now if we move all the sub elements off the parabolic surface along the direction of the received wave, we can more conveniently locate them on a plane (e.g., along the ground) provided we compensate for the extra path length by adding the appropriate delay to each sub element before adding the signals. We have now formed a “phased array” as discussed in [Sect. 7.3](#) – see, for example, Christiansen and Hogbom (1985). If the signal is monochromatic, the change in delay (which is a function of the pointing position in the sky) can be replaced by a change in phase. If the signal is not monochromatic, this change in phase will only be correct for a small region of sky around the pointing direction (called the phase center for an array); the region with no bandwidth de-correlation is called the delay beam. The signals from the sub aperture are only averaged over an area A_{tot}/n so the phased array has a much larger primary beam than the full dish. Note that the signals

formed by this phased aperture with n elements of area A_{tot}/n is identical to that of the single aperture of area A_{tot} so the S/N estimated for the single dish (Sect. 6.2) applies to a phased array with the same area.

Now consider the radiation received at a nearby point in the focal plane of the single dish or equivalently the radiation received if the pointing of the dish is changed by $\Delta\theta$. For the signals from the sub elements, this is equivalent to changing all the delays by $B \cdot \cos(\Delta\theta)$ where B is the distance from the center of the dish/array. We can now form multiple beams pointing in different directions in the sky by combining the signals with the appropriate delays.

The detected power from the sum of all the sub elements is

$$\left(\sum V_i\right)^2 = \sum (V_i)^2 + \sum (V_i \cdot V_j).$$

The first term is just the sum of self-correlations which are the total power from each element, and the second term is the sum of all possible cross-correlations between the elements. The sum of all the self-correlations includes all the emission from the sky, the atmosphere, and the ground, but these are uncorrelated between elements so do not influence the cross-correlations. Only the signal from the astronomical source will be correlated. The self-correlation term can be quite large and will be affected by gain variations in the system and variability in the radiation received from the atmosphere and the ground, making it hard to detect weak astronomical signals. It was this component that Ryle and Vonberg (1946) removed from the interferometer response when they invented the phase switch. The result is equivalent to the modern correlation interferometer array in which only the product terms are measured, either by analog or digital means.

8.4 Array Sensitivity

The previous section demonstrated the equivalence between dishes and arrays and can be used to obtain a simple sensitivity comparison. The sensitivity of the single dish is

$$\Delta S = \frac{2kT_{\text{sys}}}{A_{\text{tot}}\sqrt{\tau\Delta\nu}} = \frac{2kT_{\text{sys}}}{A_i\sqrt{n^2\tau\Delta\nu}}$$

where A_i is the area of each element. The removal of the n independent auto-correlations reduces the sensitivity of the array from n^2 independent measurements to $n(n-1)$ measurements so the sensitivity for a correlation array of total area $A_{\text{tot}} = nA_i$ is

$$\Delta S = \frac{2kT_{\text{sys}}}{A_i\sqrt{n(n-1)\tau\Delta\nu}}.$$

For a more formal derivation of the sensitivity, see, for example, Thompson et al. (2001).

Hence, we see that the sensitivity of an array approaches the sensitivity of a single dish of the same area for large n . For a point source, all cross-correlations in the array have the same S/N , so the sensitivity is independent of the element separation, and the equivalence to a single dish of the same area applies to a correlation interferometer array of any configuration. However, if the source is resolved on some baselines, the S/N will be reduced in a complex manner which will depend on the array configuration and the source structure. In this situation, it is often necessary to estimate the resulting S/N using simulations. However, a simple estimate of the approximate sensitivity to an extended source can be obtained by estimating the total area of those array elements which are close enough to not resolve the source.

9 Backends, Data Analysis, and Software

Backends, data analysis, and software are discussed in application and data reduction/analysis methods in T.L. Wilson, Volume 2, [Chap. 6](#), Sects. 9.1–9.5.

10 Types of Radio Frequency Interference (RFI) and Mitigation Strategies

At present, 1–2% of the spectrum in the meter and centimeter bands is protected for passive uses, such as radio astronomy. These regulations are coordinated by the International Telecommunications Union (ITU) and implemented by national regulations. However, future telescopes like the SKA and JVLA will have sensitivities up to 100 times greater than present sensitivities and bandwidths far exceeding the few percent covered by regulation. There are also experiments (e.g., the epoch of re-ionization, redshifted hydrogen in galaxies, or various molecular lines) which require access to arbitrary parts of the spectrum. Other experiments require very large bandwidths for sensitivity, spectropolarimetry, or spectral line information. The current regulations alone will not provide the necessary protection against RFI, so we need technology and radio quiet sites as well as regulation (Ekers and Bell 2002). See Ellingson (2005) and Kesteven (2010) for a more recent compilation of RFI mitigation strategies.

10.1 Radio Frequency Interference (RFI)

Interference may be naturally occurring or human-generated. Examples of naturally occurring interference include the following: spill-over, sun, lightning, meteors. Human-generated interference may come from broadcast services (e.g., TV, radio), voice and data communications (e.g., mobile telephones, two-way radio, wireless IT networks), navigation systems (e.g., GPS, GLONASS, Galileo), radar, remote sensing, electric fences, car ignitions, and domestic appliances (e.g., microwave ovens, Goris 1998).

Undesired interfering signals and astronomy signals can differ (be orthogonal) in a range of parameters, including frequency, time, position, polarization, distance, coding, positivity, and multipath. It is extremely rare that interfering and astronomy signals do not possess some level of orthogonality in this >8-dimensional parameter space. Signal processing systems are being developed to take advantage of the orthogonality and separate the astronomy signals from the RFI signals.

External interference may arise from fixed or moving sources. Not all methods of mitigation apply to both, and methods that work well for fixed sources may not work at all for moving sources.

10.2 RFI Mitigation Methods

There is no silver bullet for detecting weak astronomical signals in the presence of strong undesired RFI. A/D converters must be fast enough to give sufficient bandwidth, with a sufficient number of bits so that both strong and weak signals are well sampled. There are a range of

techniques that can make passive use of other bands possible, and, in general, these need to be used in a progressive or hierarchical way.

- *Remove at source* is obviously best, but may not be possible.
- *Regulation* providing radio quiet frequencies or regions.
- *Farside lobes* of primary and secondary elements must be both minimized and well characterized.
- *High dynamic range linear receivers* to allow appropriate detection of both astronomy (signals below the noise) and interfering signals (with peaks in some frequencies \gg noise).
- *Notch filters* (analog, digital, or photonic) to excise bad spectral regions; if the RFI is strong, this may require *Front-end filtering* (possibly using high-temperature super conductors) to remove strong signals as soon as they enter the signal path.
- *Clip* samples from time-based data streams to mitigate burst type interference.
- *Decoding* to remove multiplexed signals. Blanking of period or time dependent signals is a very successful but simple case of this more general approach.
- *Cancellation* of undesired signals, before correlation using adaptive filters (Barnbaum and Bradley 1998).
- *Post-correlation cancellation* of undesired signals, taking advantage of phase closure techniques (Briggs et al. 2000).
- *Parametric techniques* allow the possibility of taking advantage of known interference characteristics to excise it (Ellingson et al. 2001; Athreya 2009).
- *Adaptive beamforming* to steer spatial nulls onto interfering sources. Conceptually, this is equivalent to cancellation, but it provides a way of taking advantage of the spatial orthogonality of astronomy and interfering signals (see next section).

10.3 Adaptive Beam Nulling

This is a promising new approach and is applicable to arrays in either the aperture or focal plane, for example, Nagel et al. (2007). A physical interpretation of why you can form nulls without wrecking the synthesized beam might go as follows: Big arrays, once phased up to point in a given direction, have lots of far sidelobe nulls, which are all over the place once you get a reasonable distance from the main beam. Imagine changing the coefficients a little to get the closest null onto an interferer. Very little variation in the coefficients is required. Since the difference is so small, the main beam is hardly affected. The other nulls will shift around, of course, because they are sensitive to small changes in the coefficients.

11 General Discussion

11.1 Open Skies Policy in Radio Astronomy

Almost all radio observatories operate with an “open sky” model in which access is not limited to scientists from country or organization that operates the telescope.

This is usually justified on the basis that it guarantees the best science with the facility, whereas guaranteed access rights for scientists from funding nations favors the individual scientists more than the funding nations. The facility will still get the recognition regardless of

the nationality of the user, and with an open skies policy, it is easier to set up the large teams conducting surveys which can be made available to the entire community.

There are national benefits for financial participation in an observatory other than preferential access. These include representation in policy setting committees, involvement in future instrumental developments, and representation in time allocation committees.

11.2 Selecting the Best Telescope for Your Experiment

Given the “open skies” policy at many radio astronomy observatories, it makes sense to think about the best radio telescope for your experiment. The main decision will depend on the same telescope design criteria discussed in [Sect. 6](#). To these, you need to add geographic location and adequate sky coverage to see your source. If you need to make an image, there may be additional declination constraints, depending on the geometry of the array. For time-variable phenomena, you may also need to observe at a specific time and hence geographic longitude. Some telescopes are optimized for imaging (usually the arrays) and others for time domain astronomy (usually the single dishes).

Selecting the right frequency will depend on the scientific requirements. If specific rest frame line transitions are involved (possibly with red shifts), the observing frequency will be obvious. Note that modern correlation spectrometers and some telescope systems support more than one simultaneous frequency band, but there are usually telescope-specific constraints which still need to be considered.

Radio continuum observations are either thermal or nonthermal. If thermal, the frequency choice will depend on the optical depth with higher frequencies required for regions with significant optical depth. For nonthermal sources at low frequencies (<1 GHz), you need to consider the S/N balance between the spectrum of the sources and the background (nonthermal) noise. If polarization information is required, you will have a complex trade-off between depolarization effects favoring higher frequencies and Faraday rotation measure sensitively favoring the lower frequencies and large fractional bandwidths.

Perhaps the most critical of the other factors are the angular resolution and field of view. Angular resolution varies by a factor of 10^5 (hence 10^{10} in brightness sensitivity). Single dishes and low frequency arrays have low-resolution and high-brightness sensitivity. The current mainstream-connected arrays (VLA, GMRT, ATCA, and WSRT) have a resolution in $0''.1$ – $10''$ range and are very good when comparing with images at other wavelengths. Very high-resolution (VLBI) observations are more specialized and usually involve a very small FoV due to data transfer and processing limits.

11.3 Analog Versus Digital

This has always separated different telescope designs. In the beginning, the aperture synthesis had to wait for the computational capability to calculate Fourier transforms, and later fast Fourier transforms (Cooley and Tukey 1965). Before that, analog beamforming was the only solution. Over time, the arrays have increasingly relied on digital processing, while the single dishes used analog systems up to the final data analysis stage.

The digital vs. analog divide as a function of processing speed has evolved dramatically over time. Digital processing speed has been following Moore’s law, but most analog developments

have been much slower so the transition between analog and digital has been moving to higher frequencies and wider bandwidths. Despite these changes, the digital analog divide is still with us because the bandwidths and maximum frequencies in radio telescopes still push the maximum achievable digital signal processing limits. Currently, the digital-analog transition is in the 100 GHz range for most arrays, but for very large arrays such as SKA, the number of elements still makes fully digital difficult even at a few GHz. At high frequency, the balance shifts to larger dish size due to the increased cost contribution of the very low-noise cryogenic receivers and the cost of digital signal processing bandwidth.

In addition to these general design issues, we have also had the movement of the digital/analog transition closer and closer to the front end of the system, ultimately resulting in the digital receiver.

11.3.1 Fully Digital Receivers

In a digital receiver, the whole analog radio frequency band from the output of the low-noise amplifier is digitized directly without any prior down-conversion and analog processing. All signal processing operations are performed in the digital domain, and the digital sampler (clock) replaces the local oscillator in a conventional receiver. This is also known as a software radio. The digital receiver revolution has been dramatic and has huge impact for wireless communications. For example, see Reed (2002).

These digital receivers are extremely flexible, and production costs are low. Once designed and manufactured, they are simpler to use and replicate because there is no need to tune or match various analog components. In a digital receiver, it is also easier to implement specialized real-time signal processing techniques such as RFI suppression, programmable de-dispersion, and real-time time-domain processing.

For radio astronomy applications, the digital receivers present some special challenges which are now being overcome with modern technology. High performance and stability are required, and self-generated interference produced close to the high-gain amplifiers has to be suppressed. High bandwidths are required, but multibit samplers with up to 30 GHz clock rates are now commercially available.

11.4 General Purpose Versus Specialized Telescope Designs

This debate is still as vigorous as ever involving the cost trade-offs between specialized vs. flexible general use telescopes.

The beginning of radio astronomy provides excellent examples of discoveries made by exploring the unknown (Kellermann et al. 2009). Wilkinson et al. (2004) included a tabulation of the key discoveries in radio astronomy since the beginning of the field in 1933 to 2000.

🔗 *Figure 8-5a* plots these discoveries against time, comparing the discoveries made with special purpose instruments with those made on the larger general user facilities. It is clear that the number of discoveries made with special purpose instruments has declined with time.

🔗 *Figure 8-5b* shows that serendipitous discoveries are also more prevalent at the inception of a new branch of science.

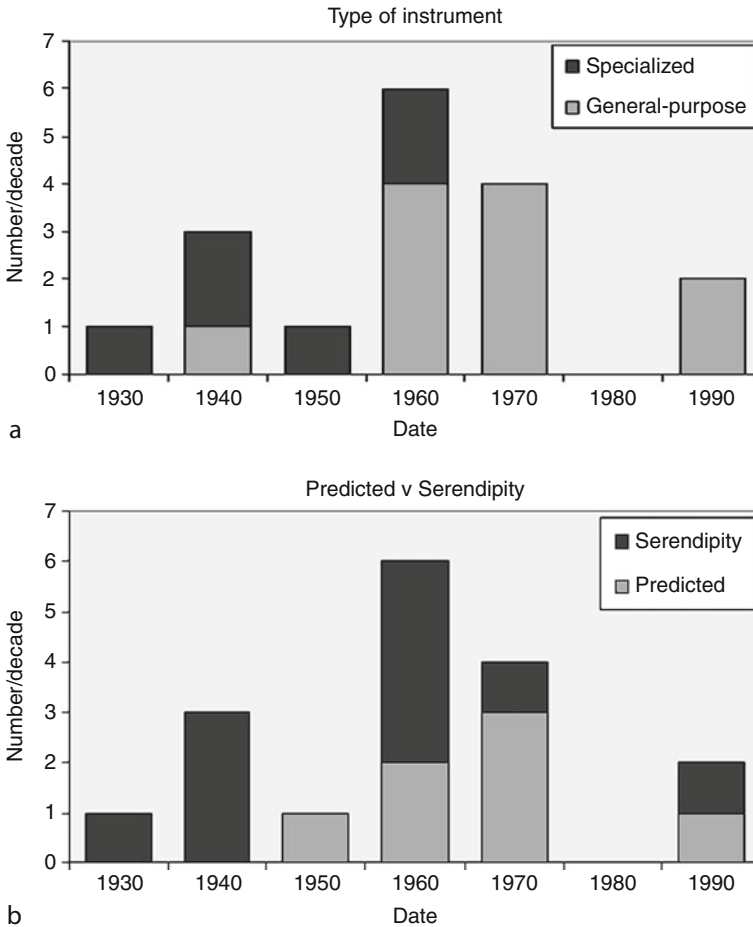


Fig. 8-5
Key discoveries in radio astronomy (From Wilkinson et al. (2004))

12 The World's Major Radio Telescopes

It is not practical to describe all the telescopes listed in Table A2, so this section is restricted to the subset of telescopes which have had the most impact based on the bibliometric analysis of Trimble and Ceja (2010).

12.1 Very Large Array (VLA, Now JVLA)

In 1965, a proposal to construct the VLA (► Fig. 8-6) was submitted to the US National Science Foundation (NSF). With twenty-seven 25-m dishes and reconfigurable baselines in a 2D Y-shaped array extending to 36 km, this was a huge step forward in sensitivity and angular resolution (Napier et al. 1983). The antennas are Cassegrain, and the receivers are in a ring at the



■ Fig. 8-6
VLA with rainbow 1985 (© Doug Johnson/Science Photo Library)

secondary focus where they can be quickly changed by a tilting 2.3-m subreflector. The 2D array provides good-quality imaging from the North Pole down to declinations of $\sim -40^\circ$. The four configurations give a wide range of resolution and brightness sensitivity. This was one of the first centimeter synthesis telescopes to provide good-quality imaging of equatorial sources.

The VLA construction commenced in 1972, and it was formally inaugurated in 1980. The VLA has been the most productive ground-based telescope ever built at any wavelength in both its number of publications and number of citations. The VLA sensitivity and imaging quality open radio wavelength observations to many fields of astronomy including stellar (mass loss rates), planets, Galactic variables, Galactic Center, and normal galaxies as well as the expected radio galaxy and quasar research. It is now undergoing a major upgrade (see ► Sect. 13.1) and will have almost complete frequency coverage with multiple low-noise dual polarization receivers from 1 to 40 GHz.

The JVLA is operated by NRAO as a National Science Foundation facility with open sky access policy.

12.2 Australia Telescope Compact Array (ATCA)

The Australian Telescope Compact Array (ATCA) started operating in 1987. It has six 22-m movable dishes and a 6-km E–W baseline. It is the premier southern hemisphere aperture synthesis telescope (Frater et al. 1992). Although it has only modest collecting area, it has multiple simultaneous frequencies and dual polarization low-noise receivers extending up to 100 GHz, very high (8 GHz) bandwidth, and low system temperature (Wilson et al. 2011).

Like the JVLA (see ► Sect. 13.1), it illustrates the degree to which technology development has enhanced radio telescope performance. All receivers are on a turret at the secondary focus which is rotated to bring the active receiver on-axis, resulting in exceptional polarization

performance. The long baselines are on an E–W rail track limiting good 2D imaging to declinations less than -20° . However, there are short (200-m) N–S baselines which provide lower-resolution 2D arrays which are especially useful for synthesis imaging at short wavelengths even near the equator. The ATCA was completed just in time to image the nonthermal radio emission of the SN1987a in the Magellanic Clouds and has continued to follow the development of the expanding shell for the last 25 years (Zanardo et al. 2010).

The ATCA is operated by CSIRO as a national facility with open sky access policy.

12.3 VLBA (Very Long Baseline Array)

The VLBA is a system of ten 25-m diameter parabolic dishes which was dedicated in 1993. With antennas distributed from Mauna Kea on the Big Island of Hawaii to St. Croix in the US Virgin Islands, the VLBA spans 8,000 km and provides the highest angular resolution of any telescope on Earth or in space.

The first and one of the very few confirmed super massive black holes was found in the galaxy NGC4258. Its mass and small size is traced by VLBA observations of a H₂O maser source in its nucleus (Miyoshi et al. 1995). In 1999, follow-up VLBA observation (Herrnstein et al. 1999) made the first direct extragalactic geometric distance measurement.

The VLBA is operated by NRAO as a US National Science Foundation facility and has had an open sky access policy.

12.4 MERLIN (Multi-element Radio Linked Interferometer Network)

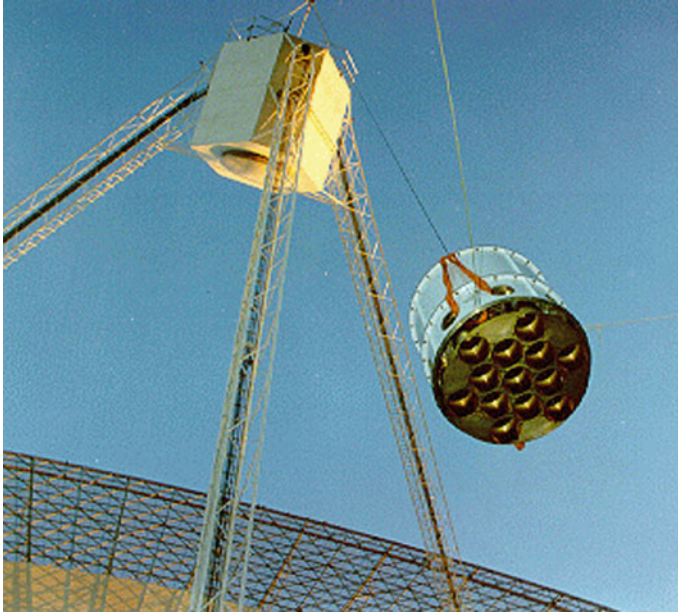
The Multi-Element Radio Linked Interferometer Network (MERLIN) is an array of radio telescopes spread across the UK. The array consists of up to seven radio telescopes and includes the Lovell Telescope, Mark II, Cambridge, Defford, Knockin, Darnhall, and Pickmere. The longest baseline is 217 km, and MERLIN operates at frequencies between 151 MHz and 24 GHz. It was originally connected in real time using microwave radio links which have now been replaced by wideband optical fiber links.

MERLIN is operated from Jodrell Bank on behalf of the Science and Technology Facilities Council as a National Facility.

12.5 Parkes

In 1960, the CSIRO Radiophysics group in Australia built a 210-ft parabolic dish now known as the Parkes 64-m radio telescope. The aerial cabin at the prime focus houses feeds and receiver equipment. The feed platform translator at the base of the aerial cabin holds up to four receivers. The translator has remotely controlled motion both up/down for focus and lateral/ rotational movement for receiver changes and polarization measurement. The alt-az mounted dish is limited in zenith angle to 59.5° .

The Parkes 21-cm Multibeam Receiver consists of a 13-beam cooled dual polarization 21-cm receiver system located at the prime focus of the 64-m dish. This receiver has had huge impact for pulsar and HI surveys and was the forerunner for multibeam receivers now available on many radio telescopes (🔗 Fig. 8-7).



■ Fig. 8-7
Installation of the multibeam receiver on the Parkes radio telescope in 1997

One of the most famous observations of the CSIRO's Parkes Observatory made soon after its completion was the Lunar Occultation of 3C273 by Hazard et al. (1963) which leads to the discovery of quasars. The occultation showed an unresolved flat-spectrum core and a 20'' steeper spectrum jet structure. The morphology and position clearly identified this strong but previously unidentified radio source with a bright 13-magnitude star with a wisp (jet) of optical emission. Schmidt (1963) obtained an optical spectrum of the star and interpreted the lines as having a redshift of 0.15.

Parkes is operated by CSIRO as a national facility with open sky access policy.

12.6 Arecibo

In 1963 at Arecibo, Puerto Rico, the US constructed the largest single aperture reflecting dish ever built. This has a 1,000-ft diameter but is a fixed spherical reflector with a movable focus. Originally designed for prime focus with a line feed, it was modified to a Gregorian with a 22-m correcting sub-reflector (Goldsmith 1996). A major component of the Arecibo telescope is the powerful radar system which enables the observatory to make radar observations of asteroids, comets, planets, and planetary satellites.

The Arecibo 1,000-ft dish was designed by Bill Gordon in the 1950s for ionospheric backscatter experiments, not for radio astronomy. It later became apparent that Gordon had overestimated the spectral width of the returned echoes in calculating the dish size needed to detect echoes from the ionosphere, and that a much smaller (and very much cheaper) dish

would be sufficient for the ionosphere experiments. However, by then, enthusiasm for a 1,000-ft dish had grown, and Gordon was able to obtain construction funds from the military who were obsessed with anything that they might learn about the ionosphere in order to detect incoming Russian missiles (Cohen, 2008, private communication), and the Arecibo telescope was built as designed (Kellermann et al. 2009).

Arecibo was operated by Cornell University since it commenced operation until 2011 when the NSF operational contract was moved to a consortium including SRI international, USRA, and University of Puerto Rico. Arecibo is a national facility, funded by the NSF with open sky access policy.

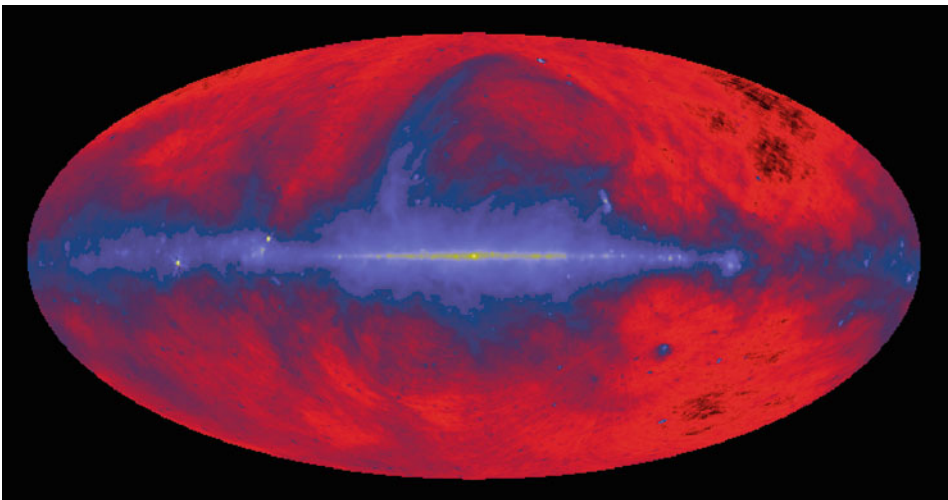
12.7 Effelsberg 100-m Telescope

The Max Planck Institute, 100-m dish in Effelsberg near Bonn, is a classical steerable parabolic dish completed in 1972. It is one of the largest fully steerable dishes in the world.

The antenna has a primary mirror 100 m in diameter and a 6.5-m secondary mirror. It operates at frequencies between 0.300 and 96 GHz. Receivers are mounted at both the primary focus and at the secondary focus of the telescope enabling rapid changes between some receivers.

The high sensitivity and good performance of the 100 m made it an excellent tool for studying the 22-GHz water vapor line both as a single dish and as the highest S/N element in VLBI observations. The first extragalactic water vapor line was found by Churchwell et al. (1977).

This very well known 408-MHz all-sky image (📍 [Fig. 8-8](#)) is from the Effelsberg 100-m telescope and the Parkes telescope (for the Southern Hemisphere). It was produced by Haslam et al. (1982) and is now the basis for the estimation of foreground nonthermal contributions to the CMB radio emission.



■ Fig. 8-8
All-sky radio emission at 408 MHz (Haslam et al. 1982)

The 100-m Effelsberg radio telescope of the Max-Planck-Institut für Radioastronomie is made available to all qualified scientists. The present policy allows the allocation of up to 40% of available observing time to visitors.

12.8 Green Bank Telescope (GBT)

The Robert C. Byrd 100-m Green Bank Telescope (GBT), replacing an older transit dish, is the last of the giant dishes to be built. It commenced operation in 2000. Unlike its predecessors, the GBT is an off-axis segment of a parabola with offset focus (both prime and Gregorian) which provides an unblocked aperture for high efficiency and minimum spectral ripple. It is one of the largest fully steerable dishes in the world.

At the same Green Bank observatory, you will now also find the original Grote Reber dish which was reassembled there by Reber in 1960, and a reconstruction of Jansky's telescope.

The GBT is operated by NRAO as a National Science Foundation facility with open sky access policy.

12.9 Westerbork Synthesis Radio Telescope (WSRT)

The forerunner of the Westerbork telescope was the Benelux Cross,¹ a joint Netherlands-Belgium project initiated by Professor Jan Oort in 1958 to use the radio astronomy source counts for cosmology.

The design was drastically modified under the influence of Jan Hogbom, a recent Ph.D. graduate from Ryle's group in Cambridge, and Chris Christiansen, from CSIRO in Sydney. The Benelux Cross was transformed into the Westerbork Synthesis Radio Telescope (WSRT) which combined aspects of aperture synthesis using movable elements and Earth rotation synthesis, from Cambridge, with the grating array concepts, from Australia. The WSRT opened in 1970 (Hogbom and Brouw 1974). In 1980, it was extended from twelve to fourteen 25-m dishes and from 1.5- to 3-km maximum E-W baseline. With its much greater sensitivity, the WSRT was able to make great advances in HI synthesis imaging and to open up areas of galactic astronomy with the observations of HII regions, interacting binaries and the Galactic Center. A phased array feed (PAF) system, called APERTIF, is being installed in the WSRT. APERTIF will cover frequencies from 1.0 to 1.7 GHz, increasing the instantaneous FoV of the WSRT to 8 deg² (Oosterloo et al. 2010).

Westerbork is operated by ASTRON, Netherlands Institute for Radio Astronomy, as an open access user facility.

12.10 Jodrell Bank

For over 50 years, the giant Lovell telescope at Jodrell Bank has been an internationally renowned landmark in the world of astronomy. It has been operating since the summer of 1957,

¹<http://www.astron.nl/radio-observatory/public/history-wsrt/benelux-cross-antenna-project/benelux-cross-antenna-project>.

just in time for the launch of Sputnik. It is a fully steerable 76-m (250-ft) parabolic antenna with receivers at the prime focus.

The Jodrell Bank 250-ft radio telescope was originally designed to detect radio echoes from cosmic ray air showers. Although this was not possible because the fast recombination in the ionized cosmic ray trail suppresses the echo below detectability, the 250-ft parabolic reflector was built in the 1950s with an upgraded surface so it could reach the 21-cm hydrogen line which was discovered in 1951. The latest upgrade took place in 2002 giving good performance at frequencies above 5 GHz.

The Lovell telescope, used as an interferometer with small telescopes, found that some radio sources had exceedingly small angular size. These were eventually identified with high-redshift stellar counterparts – the quasars.

Jodrell Bank is operated by the University of Manchester.

12.11 Giant Meterwave Radio Telescope (GMRT)

The GMRT was built near Pune in India in 1995. It is an array of 30 large fully steerable 45-m dishes and occupies a key niche for very high sensitivity at intermediate and lower radio frequencies (50 MHz–1.5 GHz). It is a 2D array with a range of baselines up to 20 km, giving an angular resolution of $1''$ at 21 cm. Although it was not in the top list of telescopes in Trimble and Ceja (2010), its impact has increased rapidly since then.

GMRT is operated by the National Center for Radio Astrophysics, a part of the Tata Institute of Fundamental Research, with open sky access policy.

13 Future Big Science Projects in Radio Astronomy

13.1 The Karl G Jansky Very Large Array (JVLA, Previously the EVLA)

After 30 years with only minimal upgrades, the VLA is undergoing a major upgrade, and the Expanded VLA is now coming into operation with 5–20 times the sensitivity, almost complete frequency coverage, and greatly enhanced spectral capability. This illustrates the dramatic impact of improved technology even though the collecting area has not changed (Perley et al. 2011).

The Expanded VLA will be known as the Karl G Jansky Very large Array (JVLA).

13.2 ALMA

This major new mm and sub-mm array is covered in [Chap. 7](#), “Submillimeter Telescopes.”

13.3 LOFAR

LOFAR is a €150-million Dutch-led project building a novel low-frequency-phased aperture arrays spread over northern Europe. It is an all-electronic telescope covering low frequencies

from 30 to 80 MHz and 120–240 MHz. The array still has some sensitivity below 30 MHz, and it may be possible to do some astronomical observations in this range. In Europe, FM radio occupancy of the band between 80 and 120 MHz makes this band unusable for radio astronomy. LOFAR will begin its operational phase in mid 2012 following a period of commissioning in 2010 and 2011.

13.4 Murchison Widefield Array (MWA)

Construction of the MWA began in February 2012. It is an all-electronic-phased array with no moving parts, observing at frequencies from 80 to 300 MHz, and located in the radio-quiet Western Australia Outback. The majority of the array tiles are concentrated into a 1.5-km core.

The MWA is an international collaboration between institutions from the US, Australia, New Zealand, and India.

13.5 Long Wavelength Array (LWA)

A phased array with plans for 53 stations each comprising 256 dipoles operating in the radio frequency range of about 20–80 MHz. The core is located at the VLA site, and baselines will eventually extend to 400 km.

The Long Wavelength Array project is a consortium led by the University of New Mexico, and includes the Los Alamos National Laboratory, the United States Naval Research Laboratories, and NASA's Jet Propulsion Laboratory.

13.6 FAST

FAST was originally conceived as an element of a future SKA proposal with a small number of large elements. It is now being developed as a single very large 500-m Arecibo-like dish. It is being built in the karst region of Guizhou province in China and uses an innovative stretched membrane structure to deform the spherical dish into a parabolic shape rather than using a line feed or a large correcting secondary mirror. It is now under construction and is expected to be completed by 2016.

13.7 SKA and the SKA Precursors

The SKA precursor facilities are being developed to demonstrate technology needed for the SKA. The two SKA precursor facilities under construction are MeerKAT and ASKAP.

13.7.1 MeerKAT

MeerKAT is a South African project to build an array of sixty-four 13.5-m diameter dishes located near Carnarvon in the Northern Cape province of South Africa. MeerKAT is part of the technology development required for the SKA. The full MeerKAT array is expected to be

ready by 2015–2016. The dishes will be equipped with a number of high-performance single pixel feeds to cover frequencies from 580 MHz up to 14 GHz.

13.7.2 ASKAP

The Australian SKA Pathfinder, ASKAP, is a project to build a telescope array of thirty-six 12-m dishes. It will be testing advanced, innovative technologies such as phased array feeds and a three axis mount to give a wide field of view (30 deg^2) with very high dynamic range (DeBoer et al. 2009).

ASKAP is being built by CSIRO at the Murchison Radio-astronomy Observatory site, an extremely low RFI environment, located near Boolardy in the Midwest region of Western Australia. All 36 antennas and their technical systems are expected to be completed in 2014.

13.7.3 SKA

The Square Kilometer Array (SKA) is a proposed radio telescope with a total collecting area of approximately one square kilometer, a frequency range from 70 MHz to 25 GHz and baselines up to at least 3,000 km from a concentrated central core. The SKA will be built in the southern hemisphere, in either South Africa or Australia. Construction of the SKA is scheduled to begin in 2016 for initial observations by 2019 and full operation by 2024.

The design will use aperture array technology for the lower frequencies and arrays of parabolic dishes at the higher frequencies. To provide a square kilometer of aperture at an acceptable cost, the SKA must make a revolutionary break with current radio telescope design. Some aspects of the technology needed are still in the development stage, and the various SKA precursors are now exploring some of the key technologies.

The construction will be a major undertaking and will be implemented in phases. Phase 1 is the initial deployment (15–20%) of the array at mid-band frequencies, Phase 2 is the full collecting area at low- and mid-band frequencies ($\sim 70 \text{ MHz}$ – 10 GHz), and Phase 3 sees the implementation at higher frequencies of 25 GHz or more.

The key science areas driving the current SKA are described in detail in “Science with the Square Kilometer Array” (Carilli and Rawlings 2004).

14 The Future

As discussed in [Sect. 2.2.2](#), the growth of radio astronomy facilities has been exponential since the beginning in 1940, but how do we maintain exponential growth? If the improvement in sensitivity has reached a ceiling, the rates of new discoveries will decline and the field will become uninteresting and slowly die. On the other hand, if we can shift to new technology or find new ways to organize our resources, the exponential increase in sensitivity can continue. Do we have such new technology to continue the exponential improvement? In radio astronomy, the combination of transistor amplifiers and their large scale integration into complex systems which can be duplicated inexpensively provides one of the keys for change. The other key technology is the computing capacity to apply digital processing at high bandwidth

thereby realizing processes such as multiple adaptive beam formation and active interference rejection in ways not previously conceivable. Finally, the move to international facilities such as the proposed SKA will also be needed to avoid the resource ceiling.

Appendix

A.1. Optical and Radio Analogs and Terminology

Radio	Optical
Antenna, dish	Telescope, element
Sidelobes	Diffraction pattern
Near sidelobes	Airy rings
Feed legs	Spider
Aperture blockage	Vignetting
Dirty beam	Point spread function (PSF)
Primary beam	Field of view
Map	Image
Source	Object
Image plane	Image plane
Aperture plane	Pupil plane
UV plane	Fourier plane
Aperture	Entrance pupil
UV coverage	Modulation transfer function
Grating responses	Aliased orders
Primary beam direction	Grating blaze angle
UV (visibility) plane	Hologram
Bandwidth smearing	Chromatic aberration
Local oscillator	Reference beam
Dynamic range	Contrast
Phased array	Beam combiner
Correlator	<i>No analog</i>
<i>No analog</i>	Correlator
Receiver	Detector
Taper	Apodise
Self-calibration	Wavefront sensing (adaptive optics)

A.2. The World's Largest Centimeter and Meter Radio Telescopes

Notes on the Table A.2:

Only operating radio telescopes with diameter greater than 25 m (or equivalent area) are included. Note that this excludes many of the important smaller mm telescopes. For simplicity, the geometric areas are given. Effective areas (see [▶ Sect. 4.2.1](#)) will be less and depend on

Country	Name	Latitude	Longitude (east)	Frequency range (GHz)	No	Size (m)	Area (sqm)	Type
Argentina	IAR	-34.88	-58.14	1.42	2	30	1,414	Dish
Australia	ATCA - Narrabri	-30.31	149.55	1.30	6	22	2,281	Dish array
Australia	Ceduna	-31.87	133.81	2.20	1	30	707	Dish
Australia	DSS43 Tidbinbilla	-35.41	148.98	1.70	1	70	3,848	Dish
Australia	Hobart	-42.81	147.44	1.40	1	26	531	Dish
Australia	MOST	-35.37	149.42	0.84	2	778 × 12	9,336	Cylinder cross
Australia	Parkes	-33.00	148.26	0.44	1	64	3,217	Dish
Canada	DRAO - synthesis telescope	49.32	-119.62	0.41	7	8.5	397	Dish array
Canada	DRAO 26-m	49.32	-119.62	1.42	1	26	531	Dish
China	Shanghai	31.10	121.20	1.62	1	25	491	Dish
China	Urumqi	43.50	87.18	0.31	1	25	491	Dish
France	IRAM - Plateau de Bure	44.63	5.91	0.08	3	15	530	Dish array
France	Nancay decametric array	47.38	2.20	0.01	144	2 × 3,500	3,526	Array
France	Nancay radio heliograph	47.38	2.20	0.15	44	5	864	Dish cross
France	Nancay radio telescope	47.38	2.20	1.30	1	200 × 40	8,000	Kraus type
Germany	Effelsberg	50.52	6.88	0.41	1	100	7,854	Dish
India	Gauribidanur	13.60	77.45	0.03	1	1,500 × 25	37,500	Array T
India	GMRT	19.10	74.05	0.38	30	45	47,713	Dish array
India	Ooty radio telescope	11.38	76.67	0.33	1	530 × 30	15,900	Cylinder
Italy	Medicina	44.52	11.65	0.32	1	32	804	Dish
Italy	Medicina northern cross	44.52	11.65	0.41		600 × 34	30,000	Cylinder T
Italy	Noto	36.88	14.99	1.33	1	32	804	Dish
Italy	Sardinia	39.50	9.24	0.41	1	64	3,217	Dish
Japan	Kashima	35.95	140.65	1.50	1	34	908	Dish
Japan	Nobeyama	35.94	138.48	1.40	1	45	1,590	Dish
Japan	Usuda	36.13	138.37	1.40	1	64	3,217	Dish
Japan	VERA	24-39	124-142	2.20	4	20	1,257	VLBI dishes
Korea	KVN	33-37	126-129	2.00	3	20	942	VLBI dishes
Mauritius	MRT	-20.13	57.73	0.15	1	2,000 × 2	25,000	Array T
Netherlands	Dwingeloo	52.81	6.40		1	25	491	Dish

Country	Name	Latitude	Longitude (east)	Frequency range (GHz)	No	Size (m)	Area (sqm)	Type
Netherlands	Westerbork	52.92	6.60	0.32	14	25	6,872	Dish array
Poland	Torun	52.91	18.56	0.15	1	32	804	Dish
Puerto Rico	Arecibo	18.34	-66.75	0.33	1	305	73,062	Fixed spherical
Russia	Kalyazin	57.22	37.90	0.61	1	64	3,217	Dish
Russia	KVAZAR	44-61	30-102	1.40	3	32	2,413	VLBI dishes
Russia	Puschino BSA	54.82	37.67	0.11	1		25,000	Phased array
Russia	Puschino DKR-1000	54.82	37.67	0.04	1	2 × 1,000	16,000	Cylinder cross
Russia	RATAN 600	43.83	41.59	1.00	1	2 × 1,812	3,624	Parabolic section
South Africa	Hartebeesthoek	-25.89	27.68	1.60	1	26	531	Dish
Spain	DSS63 Madrid	40.43	-4.25	1.70	1	70	3,848	Dish
Spain	IRAM - Pico Veleta	37.07	-3.39	1.61	1	30	707	Dish
Sweden	Onsala	57.40	11.93	1.33	1	25	491	Dish
UK	Cambridge - Ryle	52.17	0.04	0.04	8	13	1,062	Dish array
UK	Jodrell Bank - Lovell	53.24	-2.31	0.15	1	76	4,536	Dish
UK	Jodrell Bank - MkII	53.24	-2.31	0.15	1	25 × 28	550	Dish
UK	Merlin - Cambridge 32 m	52.17	0.04	0.15	1	32	804	VLBI dishes
UK	Merlin	52-54	-2 to -3	0.15	4	25	1,963	VLBI dishes
Ukraine	Evpatoriya	48.38	31.16	1.60	1	70	3,848	Dish
Ukraine	URAM 1-4	42-50	25-43	0.01	4			Interferometer
Ukraine	UTR-2 Grakovo	49.63	36.93	0.01	1	1,800 × 54 + 900 × 54	150,000	Array T
USA	DSS13 Goldstone	35.25	-116.79	2.30	1	26	531	Dish
USA	DSS14 Goldstone	35.43	-116.89	1.70	1	70	3,848	Dish
USA	Greenbank, 140'	38.44	-79.83	0.05	1	43	1,452	Dish
USA	Greenbank, interferometer	38.44	-79.83	2.10	2	26	1,062	Interferometer
USA	Haystack observatory	42.62	-71.49	2.20	1	36	1,018	Dish
USA	Owens Valley 40 m	37.00	-118.00	0.32	1	40	1,257	Dish
USA	Owens Valley interferometer	37.23	-118.29	0.50	2	27	1,145	Interferometer
USA	JVLA	34.08	-107.62	0.05	27	25	13,254	Dish array
USA	VLBA	17-48	-64 to -155	0.31	10	25	4,909	VLBI dishes

actual aperture efficiency which is a function of frequency. At best, both these will usually be 65% of the geometric area and sometimes much less. Upper and lower frequencies are based on available receivers rather than the antenna frequency range.

Groups of antennas used primarily as part of a VLBI array are not listed separately (e.g., VLBA antennas), and the range of coordinates given and number of antennas indicate the full extent of the array.

Longitude is given as an angular measurement ranging from 0° at the prime meridian to $+180^\circ$ eastward and -180° westward. For calculations, the west/east suffix is replaced by a negative sign in the western hemisphere. Confusingly, the convention of negative for east is also sometimes seen. We use the preferred convention that east be positive. Latitude south is minus.

References to Table A.2:

The most complete listings of radio telescopes and their operating frequencies are maintained by the three ITU radio astronomy spectrum management authorities:

CORF (North America) http://sites.nationalacademies.org/BPA/BPA_059065#list

CRAF (Europe) <http://www.craf.eu/raobs.htm>

RAFCAP (Asian Pacific) http://www.atnf.csiro.au/rafcap/AP_RT.htm

References

- Akgiray, A., Weinreb, S., & Imbriale, W. A. 2011, Design and measurements of a dual-polarized wideband constant-beamwidth quadruple-ridged flared horn, in IEEE Antennas and Propagation International Symposium, Spokane, Washington DC, July 2011
- Athreya, R. 2009, *AJ*, 696, 885–890
- Baars, J. W. M. 2007, The parabolic reflector antenna in radio astronomy and communications, in *Astrophysics Science library* (Heidelberg: Springer)
- Barnbaum, C., & Bradley, R. F. 1998, *AJ*, 116, 2598–2614
- Born, M., & Wolf, E. 1965, *Principles of Optics* (Oxford: Pergamon)
- Bracewell, R. N. 1986, *The Fourier Transform and Its Applications* (2nd ed.; New York: McGraw Hill)
- Brentjens, M. A., & de Bruyn, A. G. 2005, *A&A*, 441, 1217
- Briggs, F. H., Bell, J. F., & Kesteven, M. J. 2000, *AJ*, 120, 3351–3361
- Bunton, J. D. 2011, *IEEE Trans. Antenna Propag.*, 59, 2041–2046
- Burke, B. F., & Graham-Smith, F. 1996, *An Introduction to Radio Astronomy* (Cambridge, UK: Cambridge University Press)
- Carilli, C. L., & Rawlings, S. 2004, *New Astron. Rev.*, 48, 979–984
- Christiansen, W., & Warburton, J. 1955, *Aust. J. Phys.*, 8, 474–486
- Christiansen, W. N., & Hogbom, J. A. 1985, *Radiotelesopes* (2nd ed.; Cambridge, UK: Cambridge University Press)
- Churchwell, E. B., Witzel, A., Huchtmeier, W., Pauliny-Toth, I., Roland, J. & Sieber, W. 1977, *A&A*, 54, 969
- Cooley, J. W., & Tukey, J. W. 1965, *Math. Comput.* 19, 297–301
- Cordes, J. 2007 revised 2009, *Survey Metrics*. SKA Memo 109, <http://www.skatelescope.org/publications/>
- Cornwell, T. J. 1988, *A&A*, 202, 316
- Cornwell, T. J., Holdaway, M. A., & Uson, J. M. 1993, *A&A*, 271, 697
- DeBoer, D. R., Gough, R. G., Bunton, J. D., Cornwell, T. J., Beresford, R. J., Johnston, S. Feain, I. J., Schinckel, A. E., Jackson, C. A., Kesteven, M. J., Chippendale, A., Hampson, G. A., O'Sullivan, J. D., Hay, S. G., Jacka, C. E., Sweeloam, T. W., Storey, M. C., Ball, L., & Boyle, B. J. 2009, *Proc. IEEE*, 97, 1507–1521
- de Solla Price, D. J. 1963, *Little Science, Big Science* (New York, NY: Columbia University Press)
- Ekers, R. D., & Bell, J. F. 2002, Radio frequency interference, in *IAU Symposium 199: The Universe at Low Radio Frequencies*, Pune, India, 30 November–4 December 1999, ed. A. Pramesh Rao, G. Swarup, & Gopal-Krishna, 498–505

- Ekers, R. D., & Rots, A. H. 1979, in Image Formation from Coherence Functions in Astronomy, Proc. IAU Colloq. 49, Groningen, Netherlands, August 10–12, 1978, Astrophysics and Space Science Library. Vol. 76, ed. C. van Schooneveld (Dordrecht/Boston: Reidel), 61
- Ellingson, S. W. 2005, Introduction to special section on mitigation of radio frequency interference in radio astronomy. *Radio Sci.*, 40, RS5S01
- Ellingson, S. W., Bunton, J. D., & Bell, J. F. 2001, *ApJSS*, 135, 87–93
- Fisher, J. R., & Bradley, R. F. 2000, *Proc. SPIE*, 4015, 308–318. *Radio Telescopes*, ed H. R. Butcher
- Frater, R. H., Brooks, J. W., & Whiteoak, J. B. 1992, *J. Electr. Electron. Eng. Aust.*, 12, 103–12
- Goldsmith, P. F. ed. 1988, in *Instrumentation and Techniques for Radio Astronomy* (New York: IEEE)
- Goldsmith, P. F. 1994, *Quasioptical Systems: Gaussian Beam Quasioptical Propagation and Applications* (New York: Wiley/IEEE)
- Goldsmith, P. F. 1996, *IEEE*, 15, 38–43
- Goris, M. 1998, *Categories of Radio Interference*, NFRA Technical Report – 415/MG/V2.3
- Harwit, M. 1981, *Cosmic Discovery – The Search, Scope and Heritage of Astronomy* (New York, NY: Basic Books)
- Haslam, C. G. T., Salter, C. J., Stoffel, H., & Wilson, W. E. 1982, *A&A Suppl* 47, 1
- Hay, S., O’Sullivan, J., & Mitra, R. 2008, *IEEE Antennas and Propagation Society International Symposium, AP-S*, San Diego, CA, July 5–11 2008, 1–4
- Hazard, C., Mackey, M. B., & Shimmins, A. J. 1963, *Nature*, 197, 1037–1039
- Herrnstein, J. R., Moran, J. M., Greenhill, L. J., Diamond, P. J., Inoue, M., Nakai, N., Miyoshi, M., Henkel, C., & Riess, A. 1999, *Nature*, 400, 539–541
- Hogbom, J. A., & Brouw, W. N. 1974, *A&A*, 33, 289
- Jansky, K. G. 1933a, *Nature*, 132, 66
- Jansky, K. G. 1933b, *Proc. IRE*, 21, 1387–1398
- Jenkins, F. A., & White, H. E. 2001, in *Fundamentals of Optics* (4th ed.; New York: McGraw-Hill)
- Kawaguchi, N., Sasao, T., & Manabe, S. 2000, in *Radio Telescopes*, ed H. Butcher, *Proc. SPIE*, Vol. 4015 (Washington, DC: SPIE), 544–551
- Kellermann, K. I., Cordes, J. M., Ekers, R. D., Lazio, J., & Wilkinson, P. N. 2009, in *Accelerating the Rate of Astronomical Discovery – SPSS*, IAU GA, Rio de Janeiro, Brazil, August 11–14
- Kellermann, K. I., & Moran, J. M. 2001, *Ann. Rev. A&A*, 39, 457–509
- Kesteven, M. 2010, *Proc. RFI Mitig. Workshop*, 29–31 March 2010. Groningen, the Netherlands. PoS(RFI2010)007
- Kildal, P.-S., Jian Yang, Karandikar, Y., Wade-falk, N., Pantaleev, M., & Helldner, L. 2009, Development of a coolable 2–14 GHz Eleven feed for future radio telescopes for SKA and VLBI 2010, in *Electromagnetics in Advanced Applications*, 2009. ICEAA ’09, Turin, Italy, 545–547
- Kraus, J. D. 1986, *Radio Astronomy* (2nd ed.; Powell, OH: Cygnus-Quasar)
- Large, M. I., Campbell-Wilson, D., Cram, L. E., Davison, R. G., & Robertson, J. G. 1994, *Astron. Soc. Aust. Proc.* 11, 44–49
- Livingston, M. S., & Blewett, P. 1962, *Particle Accelerators* (New York: McGraw Hill)
- Love, A. W. ed. 1976, in *Electromagnetic Horn Antennas* (New York: IEEE)
- Machin, K. E. 1951, *Nature*, 167, 889–89
- Macquart, J.-P., Ekers, R. D., Feain, I., & Johnston-Hollitt, M. 2012, *ApJ*, 750, 15
- McCready, L. L., Pawsey, J. L., & Payne-Scott, R. 1947, *Proc. R. Soc. A*, 190, 357–375
- Mills, B. Y., & Little, A. G. 1953, *Aust. J. Phys.*, 6, 272
- Miyoshi, M., Moran, J., Herrnstein, J., Greenhill, L., Nakai, N., Diamond, P., & Inoue, M. 1995, *Nature*, 373, 127–129
- Mollick E. 2006, *Establishing Moore’s law*. *IEEE Ann. History Comput.*, 28(3), 62–75
- Moore G. E. 1965, *Cramming more components onto integrated circuit*. *Electronics*, 38, 8
- Nagel, J. R., Warnick, K. F., Jeffs, B. D., Fisher, J. R. & Bradley, R. 2007, *Radio Sci.*, 42, RS6013
- Napier, P. J., Thompson, A. R., & Ekers, R. D. 1983, *Proc. IEEE*, 71, 1295–1320
- O’Brien, P. A. 1953, *MNRAS*, 113, 597–612
- Oosterloo, T., Verheijen, M., & van Cappellen, W. 2010, in *Proc. ISKAF2010 Sci. Meet.*, June 10–14 2010, Assen, the Netherlands, PoS(ISKAF2010) 043
- Perley, R. A., Chandler, C. J., Butler, B. J., & Wrobel, J. M. 2011, *ApJL*, 739, L1
- Radhakrishnan, V., 1990, in *Modern Radio Science*, from URSI General Assembly, Prague, Czechoslovakia, ed J. Bach Anderson (Published for the International Union of radio Science and the ICSU Press by Oxford University Press, UK), 187
- Radhakrishnan, V. 1999, in *Synthesis Imaging in Radio Astronomy II*, A Collection of Lectures from the Sixth NRAO/NMIMT Synthesis Imaging Summer School, ASP Conf. Ser. 180, ed. G. B. Taylor, C. L. Carilli, & R. A. Perley (San Francisco, CA: ASP), 671
- Reber, G. 1958, *Proc. IRE*, 46, 15–23

- Reed, J. H. 2002, *Software Radio: A Modern Approach to Radio Engineering* (Upper Saddle River, NJ: Prentice Hall)
- Reich, W., Kalberla, P., Reif, K., & Neidhöfer, J. 1978, *A&A*, 76, 92
- Riesselmann, K. 2009, *Deconstruction: Livingston plot. Symmetry*, 6, 30
- Ryle, M. 1952, *Proc. R. Soc. Lond. A*, 211(1106), 351–375
- Ryle, M., & Hewish, A. 1960, *MNRAS*, 120, 220
- Ryle, M., & Neville, A. C. 1962, *MNRAS*, 125, 39
- Ryle, M., & Vonberg, D. D. 1946, *Nature*, 158, 339–340
- Schmidt, M. 1963, *Nature*, 197, 1040–1041
- Stanier, H. M. 1950, *Nature*, 165, 354–355
- Sullivan, W. T. 2009, *Cosmic Noise* (New York, NY: Cambridge University Press)
- Taylor, G. B., Carilli, C. L., & Perley, R. A. 1999, *Synthesis Imaging in Radio Astronomy II, A Collection of Lectures from the Sixth NRAO/NMIMT Synthesis Imaging Summer School. ASP Conf. Ser. 180* (San Francisco, CA: ASP)
- Tegmark, M., & Zaldarriaga, M. 2010, *Phys. Rev. D*, 82, id. 103501
- Thompson, A. R., Moran, J. M., & Swenson, G. W. 2001, *Interferometry and Synthesis in Radio Astronomy* (New York, NY: Wiley)
- Trimble, V., & Ceja, J. A. 2010, *Astron. Nachr.*, 331, 338
- Wild, J. P. 1967, *Proc. Instn. Radio Electron. Engrs. Aust.*, 28(9), 279–291
- Wilkinson, P. N., Kellermann, K. I., Ekers, R. D., Cordes, J. M., & Lazio, T. J. W. 2004, *The exploration of the unknown. New Astron. Rev.*, 48(11–12), 1551–1563
- Wilson, T. L., Rohlfs, K., & Hüttemeister, S. 2008, *Tools of Radio Astronomy* (5th edn.; Heidelberg: Springer)
- Wilson, W. E., Ferris, R. H., Axtens, P., Brown, A., Davis, E., Hampson, G., Leach, M., Roberts, P., Saunders, S., Koribalski, B. S., et al., 2011, *MNRAS*, 416, 832–56
- Zanardo, G., Staveley-Smith, L., Ball, L., Gaensler, B. M., Kesteven, M. J., Manchester, R. N., Ng, C.-Y., Tzioumis, A. K., & Potter, T. M. 2010, *ApJ*, 710, 1515–1529

9 Space Telescopes in the Ultraviolet, Optical, and Infrared (UV/O/IR)

*Erin Elliott · Matt Mountain · Marc Postman ·
Anton Koekemoer · Leonardo Ubeda · Mario Livio*
Space Telescope Science Institute, Baltimore, MD, USA

1	<i>Introduction: Advantages of Observatories in Space</i>	365
1.1	Atmospheric Extinction	366
1.2	Atmospheric Emission and Background	367
1.3	Atmospheric Turbulence and a Resolution Comparison	369
1.4	A Sensitivity Comparison	370
1.4.1	Method and Assumptions	373
1.5	Comparison of Exposure Times when Imaging Faint Objects	374
1.6	Very High-Contrast Imaging	379
2	<i>Space Observatories in the UV/O/IR</i>	380
2.1	Space Observatory Elements	381
2.2	Hubble Space Telescope (HST)	383
2.3	Spitzer Space Telescope (SST)	385
2.4	James Webb Space Telescope (JWST)	388
2.5	Advanced Technology Large-Aperture Space Telescope (ATLAST) and Other Future Missions	390
3	<i>Science Goals, Cost, and Productivity</i>	391
3.1	Early Design Trades	392
3.2	Mission Cost	394
3.3	Mission Productivity Metrics	394
4	<i>Orbits</i>	396
4.1	Low Earth Orbit	396
4.2	Drift-Away Orbit	398
4.3	Libration Point Orbit at Sun-Earth L2	399
4.4	Orbits and Communications for Future Missions	400
5	<i>Packaging and Launch Vehicles</i>	400
6	<i>Optical Considerations</i>	404
6.1	Optical Designs and FOV Allocations	405
6.2	Stray Light Control	406

6.3	Lightweight Mirror Technologies	408
6.4	Optical Alignment	410
6.5	Optical Alignment for Future Missions	413
7	<i>Pointing and Control Systems</i>	414
7.1	Pointing and Control Systems for Future Missions	418
8	<i>Thermal Systems</i>	419
8.1	Basic Components of Thermal Control Systems	420
8.2	Radiative and Conductive Heat Transfer	421
8.3	HST Thermal Design	422
8.4	SST Thermal Design	423
8.5	JWST Thermal Design	425
8.6	The Future in Thermal Designs	426
9	<i>Conclusion</i>	426
	<i>Acknowledgments</i>	427
	<i>References</i>	428

Abstract: Space telescopes are essential for advancing our understanding of the physics of the cosmos. The vacuum environment of space means the light entering the aperture of a space telescope has not suffered any atmospheric extinction, enabling observations in ultraviolet, far-infrared, and X-ray wavelengths. At wavelengths that are accessible from both space and ground-based facilities, the sky background levels in space are considerably lower than those at ground-based sites, particularly in the near and mid-infrared. The lack of atmospheric turbulence allows space telescopes to obtain diffraction-limited image quality without the need for adaptive optics. Their inherently high stability, and ability to observe for long continuous periods of time, allows observations to be performed from space that are difficult or impossible to accomplish from the ground. In [▶ Sect. 1](#), comparisons are made between the sensitivities, spatial resolutions, and exposure times for ground- and space-based observatories.

Fundamentally, the science objectives drive the design of a space telescope, but these have to be balanced by the technological maturity of the space systems needed, the availability of suitable launch vehicles, and ultimately the overall mission cost. Today, it is normal practice to use cost models to attempt to capture the physical parameters that drive cost. If the early estimates of mission cost become too large, science objectives may have to be reduced, or new technologies developed and the mission redesigned accordingly. In [▶ Sect. 3](#), the types of high-level trades that have to be made in designing and building a space observatory are discussed.

To illustrate the kind of trades that are undertaken, three space telescopes that have different science objectives and span nearly 30 years in launch dates are described: the Hubble Space Telescope (HST), Spitzer Space Telescope (SST), and the James Webb Space Telescope (JWST). Overviews of the three missions are given in [▶ Sect. 2](#). HST is a warm observatory with a 2.4 m primary mirror diameter in low Earth orbit launched in 1990. It observes in ultraviolet, optical, and near-infrared (NIR) wavelengths. SST is a cold infrared observatory with a 0.85 m primary mirror in a drift-away orbit launched in 2003. JWST will be a cold, infrared observatory with a 6.6 m primary mirror in orbit at the second Lagrange point of the Earth-Sun system (L₂). JWST is currently scheduled to launch in 2018.

Different orbits were selected for HST, SST, and JWST, as discussed in [▶ Sect. 4](#). A low Earth orbit was chosen for HST to allow for human servicing that changed and improved the observatory over its multi-decade lifetime. SST's heliocentric drift-away orbit was selected because it minimized the mission mass and allowed the observatory to continually radiate heat into deep space, but contact with the mission will eventually be lost. JWST's orbit about the metastable Sun-Earth Lagrange Point 2 (L₂) keeps the Sun, Earth, and Moon constantly behind its sophisticated sunshield, enabling the telescope assembly to passively cool to 40 K, operate in a highly stable thermal environment, and stay in constant communication with the ground system for its 10-year mission using NASA's Deep Space Network (DSN).

The ratios of telescope aperture diameter to launch volume and mass have increased over time, as discussed in [▶ Sect. 5](#), driven by the need for missions to fit into available launch vehicles. For HST and JWST, some components are folded to fit into the launch vehicles and are deployed after launch. JWST, in addition, requires that optical alignment be carried out on orbit because the optics themselves deploy after launch ([▶ Sect. 6.4](#)).

Optical trades are discussed in [▶ Sect. 6](#). HST and SST are compact, two-mirror Cassegrain telescopes. HST minimizes the number of mirrors in order to achieve reasonable throughput at ultraviolet wavelengths. JWST has a three-mirror design that achieves a reasonable field of view to accommodate multiple instruments despite the very large primary mirror diameter. The three-mirror design also produces an accessible pupil, and a planar mirror and mask at

that location are used for fine-pointing adjustments and stray light control. Because it operates in the IR, JWST's throughput remains high despite the additional reflections in the telescope.

HST uses a glass primary with a low thermal expansion, which enabled it to be polished to reach the lower surface errors required for observing at visible and ultraviolet wavelengths. JWST and SST, operating in infrared wavelengths, were able to make use of a beryllium primary mirror technology with much lower areal density than that of HST's light-weighted glass primary mirror. The beryllium technology has excellent cryogenic properties as discussed in [Sect. 6.3](#).

HST's telescope and instruments operate near room temperature, with the exception of its actively cooled IR and CCD detector systems. HST experiences a somewhat unstable thermal environment because of its low Earth orbit, as discussed in [Sect. 8](#). SST and JWST must operate cold to ensure that the sensitivity of the infrared detectors is limited by the shot noise from the deep-space background rather than by infrared emission from the telescope structure. SST uses a combination of passive cooling for the optics and active cooling for the detectors and instruments in order to reach their respective operational temperatures in its drift-away orbit. JWST relies almost solely on passive cooling in its orbit about the Sun-Earth L2 point, with the exception of the detectors on the mid-infrared instrument which are actively cooled by a mechanical cooler.

Many of the science objectives for future space telescopes will build on the science legacies of the telescopes covered in this chapter and will require larger primary mirror diameters. With a combination of greater angular resolution and higher sensitivity, future observatories will be capable of detecting signs of life on extrasolar planets, will reveal the mechanisms that control early galaxy formation, and will provide windows into new and unexplored regimes of the universe from ultraviolet to infrared wavelengths. Such missions are likely to make use of new, larger launch vehicles, but will still need to fold more compactly for launch and have lower areal densities than current observatories. Consequently, automatic deployments or in-space assembly will be required, as will on orbit optical alignment. Future large telescopes will likely also incorporate real-time active and adaptive optical and alignment control; such technology has been successfully developed for the recent generation of very large ground-based telescopes. With its thermal stability and high observing efficiencies, L2 will continue to be an attractive orbit for these future missions. Other orbits might be selected if there is renewed interest in servicing these future "Great Observatories."

Keywords: Active, Advanced mission cost model, Alignment, Ariane, ATLAST, Background, Cassegrain, Communication, Conductive, Cost, Cost model, Delta, Deployment, Design trade, Drift-away, DSN, Exposure times, Extinction, Fairing, Field of view, Fine guiders, Fine pointing, FWHM, Gyroscopes, Heat transfer, High-contrast, Hubble, Instrument, James Webb, L2, Lagrange point, Launch vehicle, LEO, Lightweight mirrors, MLI, Orbit, OTA, Passive, Payload, Pointing and control, Primary mirror, Productivity, PSF, Pupil, Radiative, Reaction wheels, Resolving power, Ritchey-Chretien, Science system engineering, Sensitivity, Shroud, Signal-to-noise, Space launch system, Spacecraft bus, SST, Star trackers, Stray light, Sunshield, TDRSS, Thermal stability, Thrusters, TMA, Wavefront sensing

List of Abbreviations: ACS, Attitude Control System; ACS, Advanced Camera for Surveys (HST Instrument); AMCM, NASA Advanced Mission Cost Model; AO, adaptive optics; AOS, Aft Optical System (JWST component); AS, aft shroud (HST component); ATLAST, Advanced Technology Large-Aperture Space Telescope; AU, astronomical unit; BATC, Ball Aerospace and

Technologies Corporation; *CCD*, charge-coupled device; *COBE*, Cosmic Background Explorer; *COS*, Cosmic Origins Spectrograph (HST instrument); *CSA*, Canadian Space Agency; *CTA*, Cryogenic Telescope Assembly (SST component); *DOF*, degree of freedom; *DSN*, Deep Space Network; *ESA*, European Space Agency; *FGS*, Fine Guidance Sensors; *FOSR*, Flexible Optical Solar Reflector; *FOV*, field of view; *FS*, forward shell (HST component); *FSM*, Fine Steering mirror (JWST component); *FWHM*, full width half maximum; *GALEX*, Galaxy Evolution Explorer; *GMOS*, Gemini Multi-object Spectrographs (Gemini instrument); *GSFC*, Goddard Space Flight Center; *HST*, Hubble Space Telescope; *ICRP*, Independent Comprehensive Review Panel; *IFOV*, instantaneous field of view; *IR*, infrared; *IRAC*, Infrared Array Camera (SST instrument); *IRAS*, Infrared Astronomical Satellite; *IRS*, Infrared Spectrograph (SST instrument); *ISIM*, Integrated Science Instrument Module (JWST component); *ISO*, Infrared Space Observatory; *JPL*, Jet Propulsion Laboratory; *JWST*, James Webb Space Telescope; *LEO*, low-Earth orbit; *LST*, Large Space Telescope (now HST); *MIC*, Multi-Instrument Chamber (SST component); *MIMF*, Multi-Instrument Multifield; *MIPS*, Multiband Imaging Photometer for SIRTf (SST instrument); *MIRI*, Mid-Infrared Spectrometer (JWST instrument); *MLI*, multi-layer insulation; *NASA*, National Aeronautics and Space Administration; *NICMOS*, Near Infrared Camera and Multi-Object Spectrometer (HST instrument); *NIR*, near infrared; *NIRCam*, Near-Infrared Camera (JWST instrument); *NIRI*, Near InfraRed Imager (Gemini instrument); *NIRSpec*, Near-Infrared Spectrometer (JWST instrument); *NIRISS*, Near-infrared Imaging Slitless Spectrometer; *OTA*, Optical Telescope Assembly; *OTE*, Optical Telescope Element (JWST component); *PCRS*, Pointing Control Reference Sensor (SST component); *PMT*, photomultiplier tube; *PSF*, point spread function; *PWFS*, Peripheral wave front sensor (Gemini equipment); *QE*, quantum efficiency; *RMS*, root mean square; *r.p.m.*, rotations per minute; *ROC*, radius of curvature; *S/N*, signal-to-noise ratio; *SAA*, South Atlantic Anomaly; *SAFIR*, Single-Aperture Far-Infrared Observatory; *SIRTf*, Space Infrared Telescope Facility; *SLS*, Space Launch System; *SPICA*, Space Infrared Telescope for Cosmology and Astrophysics; *SST*, Spitzer Space Telescope; *STIS*, Space Telescope Imaging Spectrograph (HST instrument); *STS*, Space Transportation System (Space Shuttles); *TDRSS*, NASA's Tracking and Data Relay Satellite System; *TEC*, thermoelectric cooler; *THEIA*, Telescope for Habitable Exoplanets and Interstellar/Intergalactic Astronomy; *TMA*, three-mirror anastigmat; *UDF*, ultra-deep field; *ULE*, ultra-low expansion (glass); *UV*, ultraviolet; *UV/O/IR*, ultraviolet/optical/infrared; *VLT*, Very Large Telescope (ground-based); *WFE*, wavefront error; *WFC3*, Wide-field Camera 3 (HST instrument); *WFSC*, Wavefront Sensing and Control (alignment process for JWST); *WISE*, Wide-Field Infrared Survey Explorer; *WMAP*, Wilkinson Microwave Anisotropy Probe; *XMM*, X-ray Multi-mirror Mission (ESA)

1 Introduction: Advantages of Observatories in Space

In all areas of astronomy and astrophysics, forefront research requires obtaining data from both space and ground-based observatories. On the ground, very large aperture diameters can be constructed for a cost per unit collecting area that is much lower than that of space-based observatories. Ground-based observatories are also easily upgraded. However, ground-based observatories are subject to key limitations tied principally to the physics of the atmosphere. Atmospheric extinction – the absorption and scattering of radiation from space due to the atomic, molecular, and aerosol components of the atmosphere – limit observations from the

ground to specific windows in the optical, near-infrared, sub-millimeter, and radio wavelengths. At certain wavelengths, the emission from the atmosphere itself becomes significant, and shot noise on this atmospheric emission becomes a significant noise source for faint astronomical observations. In addition, because ground-based telescopes sit at ambient atmospheric temperature (typically 0°C at good sites), the thermal emission from the optics, telescope structure, and atmosphere can overwhelm the signals from faint astrophysical sources at wavelengths in the thermal infrared, beyond about 2 μm. Finally, atmospheric turbulence can limit the sharpness of the system's point spread function (PSF), which is the width of an image of a perfect point source (unresolved star). For certain observations, turbulence also limits the dynamic range of measurements from the ground. Adaptive optics (AO) systems can partially compensate for the atmospheric turbulence encountered by ground-based observatories but are also subject to limitations on the fields of view, wavelength, and the availability of sufficiently bright reference stars.

Assuming a background-limited observation in which the detectors are not a dominant source of noise, a simple formulation of the achievable signal-to-noise ratio (S/N) for a telescope observing a small, faint astronomical source over a given time is given by

$$\frac{\text{signal}}{\text{noise}} \propto \frac{\text{telescope_diameter}}{\text{image_size}} \cdot \sqrt{\frac{QE_\lambda}{B_\lambda}}, \quad (9.1)$$

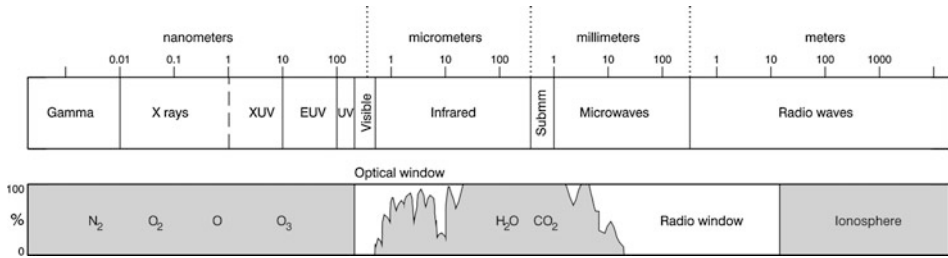
where QE_λ is the total system quantum efficiency, which is relatively high for modern ground- and space-based systems operating in optical and infrared wavelengths. However, atmospheric absorption features introduce greater wavelength dependence in the QE_λ for ground-based systems. B_λ represents the sky and telescope background and is generally low for space telescopes. The delivered image quality is characterized by the width of the delivered PSF (image_size).

Without atmospheric turbulence, the PSF width is limited only by the quality of the telescope's optics and the stability of the system. Space-based observatories can produce diffraction-limited images over a wide range of wavelengths and over wide fields of view. As [9.1](#) shows, space-based systems, with their combination of low background and better image quality, have the potential to reach significantly higher sensitivities than ground-based systems with equivalent collecting areas.

1.1 Atmospheric Extinction

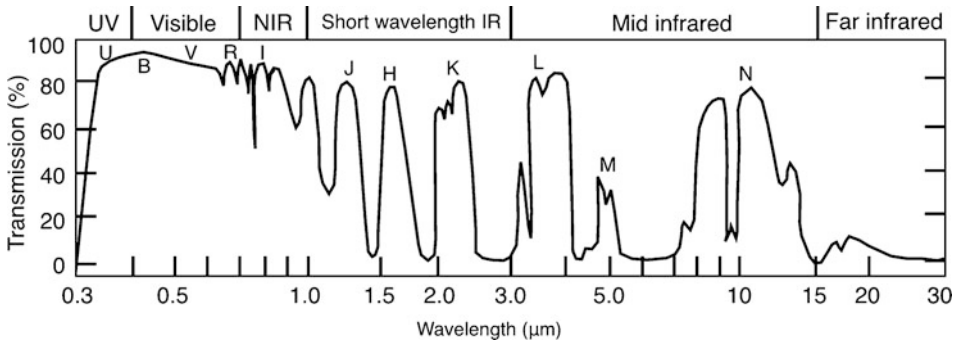
Atmospheric extinction reduces the flux of an extraterrestrial source through absorption and scattering of photons by the atmosphere. Total atmospheric extinction as a function of wavelength is shown in [Fig. 9-1](#). Ground-based observations are not possible for wavelengths where the atmospheric extinction is high: below approximately 300 nm, from 30 μm up to 1 mm, and beyond 10 m.

Ground-based observations are possible in the transmission “windows” that occur in the visible, the near UV, the near-infrared, and the radio wavelength ranges. A detail of atmospheric transmission in visible and infrared wavelengths is shown in [Fig. 9-2](#).



■ Fig. 9-1

The electromagnetic spectrum and the absorption of the atmosphere as a function of wavelength, with an indication of the primary absorbing molecules (After Bely 2003)



■ Fig. 9-2

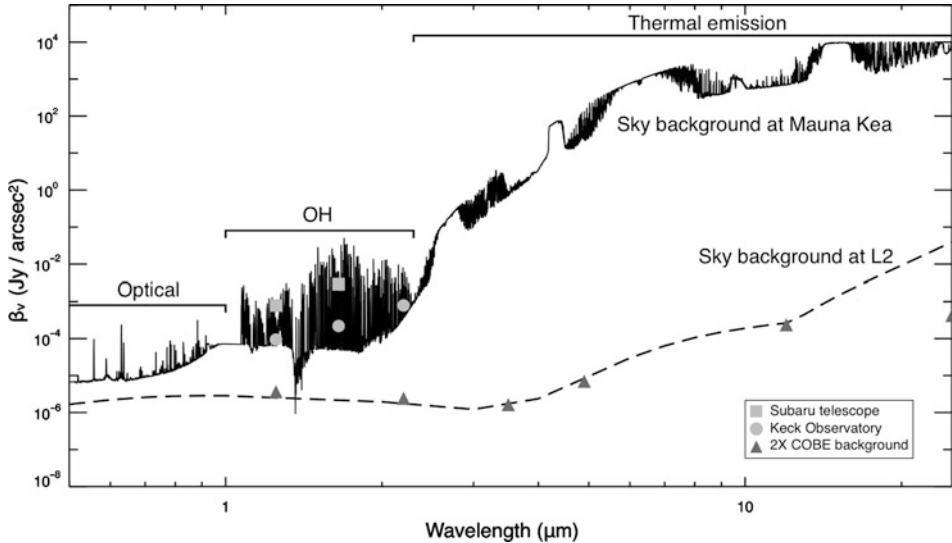
Atmospheric transmission in wavelengths from the ultraviolet to the far infrared. The letters denote the names of several standard-wide passbands that are defined by regions of high atmospheric transmission (After Elachi 1987)

1.2 Atmospheric Emission and Background

For astronomical observations, atmospheric emission introduces unwanted background noise (the shot noise on the detected background photons) to an image or spectrum measured by a ground-based telescope, reducing the signal-to-noise ratio of the measurement given by (9.1). A comparison of the typical background emission for a good, high-altitude ground site and the minimum sky background attainable from space is shown in Fig. 9-3. The black curve is the typical optical/infrared (O/IR) background for a ground-based telescope located at a high-altitude site like Mauna Kea, Hawaii.¹

The gray dashed curve in Fig. 9-3 represents an estimate of the sky background for JWST operated at 40 K and located at L2. Below 13 micrometers, the background is conservatively estimated as twice the cosmic background as measured by the Cosmic Background Explorer (COBE)

¹<http://www.gemini.edu/sciops/telescopes-and-sites/observing-condition-constraints/optical-sky-background>



■ Fig. 9-3

Typical “dark time” visible and infrared background emission for a ground-based telescope as measured at Mauna Kea (Hawaii). An estimate of the background for a cryogenically cooled telescope at L2 is shown for comparison (After Gillet and Mountain 1998)

(Gillet and Mountain 1998). Outside the Earth’s atmosphere, the sky background is dominated by the superposition of many partially or completely unresolved discrete sources of emission, including zodiacal light, light from Galactic and extragalactic objects, and emission from dust outside the solar system. Thermal emission from the observatory itself can also be a nontrivial component of background noise. For infrared telescopes in space, the optics, instruments, and detectors must be cooled to prevent the observatory from emitting in the wavelengths of interest. Beyond about 13 μm , JWST’s background curve is dominated by thermal emission from the 40 K telescope mirrors and structure.

In the optical regime, the background seen by space-based telescopes is only about three to ten times lower than that seen by ground-based systems. At night, atmospheric emission in the near infrared is dominated by fluorescence from OH^- molecules. Ground-based observations using high-resolution spectroscopy can detect signals from within the OH background (essentially observing between the OH lines) if the spectrometer’s resolving power is high. Resolving power (R) is a measure of a spectrometer’s ability to resolve features in an image that is spectrally dispersed over the detector:

$$R = \lambda / \Delta\lambda, \quad (9.2)$$

where $\Delta\lambda$ is the spectral resolution, which is the smallest difference in wavelengths that can be distinguished (or resolved) in the dispersed spectrum at a wavelength of λ . For $R > 5,000$, space-based observations have significantly less advantage over the ground in terms of background.

For broadband imaging ($R \sim 5$), however, not only is the ground-based background noise dominated by the shot noise on the integrated OH emission but the “OH airglow” fluctuates in a non-Gaussian way, limiting the accuracy of ground-based infrared observations from about

1 to 2.3 μm (Ramsay et al. 1992). In addition during periods called “bright time,” light from the Moon that scatters off the atmospheric aerosols can also increase the visible and near-infrared backgrounds.

Beyond 2.3 μm , the ground-based background is dominated by thermal emission from the telescope itself, with temperatures of about 230–280 K, and the thermal emission from molecules in the lower atmosphere with an effective temperature of approximately 250 K. This results in an exponential rise of background flux with wavelength that dramatically reduces a ground-based telescope’s sensitivity at those wavelengths.

1.3 Atmospheric Turbulence and a Resolution Comparison

If an observatory is “diffraction-limited,” the full width at half maximum (FWHM) of its PSF at a given wavelength, λ , is

$$FWHM \text{ (arcsec)} \sim 0.2\lambda(\mu\text{m})/D(\text{m}), \quad (9.3)$$

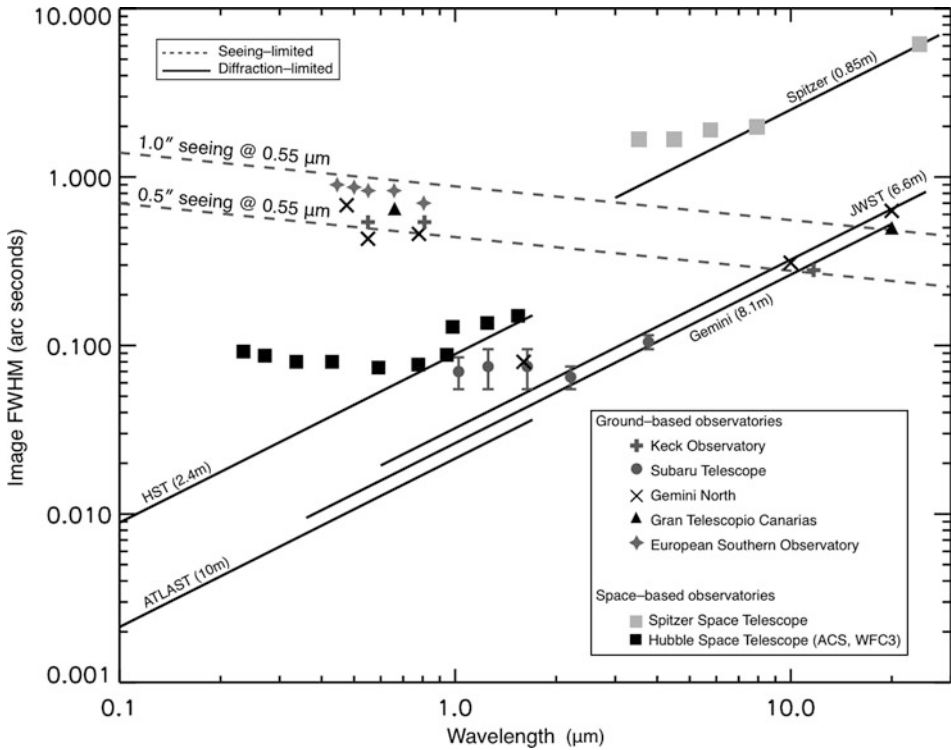
where the *FWHM* has units of arcseconds, *D* is the effective diameter of the observatory’s primary mirror in meters, and λ has units of microns. If the delivered PSF width is limited only by the fundamental physics that govern the diffraction of light, rather than by other effects such as mirror surface or alignment errors, the system is referred to as “diffraction-limited” (Schroeder 1987). Well-figured and well-controlled space-based systems can be designed and built to achieve diffraction-limited imaging over a wide range of wavelengths and over wide fields of view.

For a ground-based telescope without AO, the PSF size is dominated by atmospheric turbulence rather than by the diameter of the primary mirror. This is true even if the telescope optics are designed to be diffraction-limited. Typically, a coherence or Fried length, R_o , is used to characterize the turbulence at a good astronomical ground-based site, and the resulting time-averaged image *FWHM* can be estimated using (9.3) with R_o replacing *D*. At visible wavelengths, where R_o can be as large as 30 cm at good sites, the PSF width, or “seeing,” is about 0.37 arcsec for a wavelength of 0.55 μm .

R_o is proportional to $\lambda^{6/5}$ so the seeing improves as $\lambda^{-1/5}$ as the wavelength increases. Today, AO technologies enable telescopes as large as 10 m to compensate for this turbulence beyond wavelengths of about 1.2 μm , but only over restricted fields of view (typically tens of arcseconds to a maximum of a few arcminutes with newer tomographic AO systems).

Figure 9-4 shows the dependence of an image’s *FWHM* on wavelength for different optical configurations both on the ground and in space. The solid lines represent theoretical diffraction limits for five telescopes and are provided over the range of their operational wavelengths: Spitzer Space Telescope ($D = 0.85$ m), Hubble Space Telescope ($D = 2.4$ m), James Webb Space Telescope ($D = 6.6$ m), Gemini North ($D = 8.1$ m), and a proposed space-based mission, the Advanced Technology Large Aperture Space Telescope (ATLAST) ($D = 10$ m) (Postman 2009).

For comparison, two seeing-limited curves are included; the first shows 1.0 arcsec seeing at 0.55 μm and the other shows a median seeing case of 0.5 arcsec at the same wavelength. Median seeing and diffraction limits become comparable at a wavelength of about 10 μm for the Gemini observatory. This indicates that at wavelengths less than 10 μm , the use of AO becomes mandatory in order to make full use of the image quality from a very large ground-based telescope.



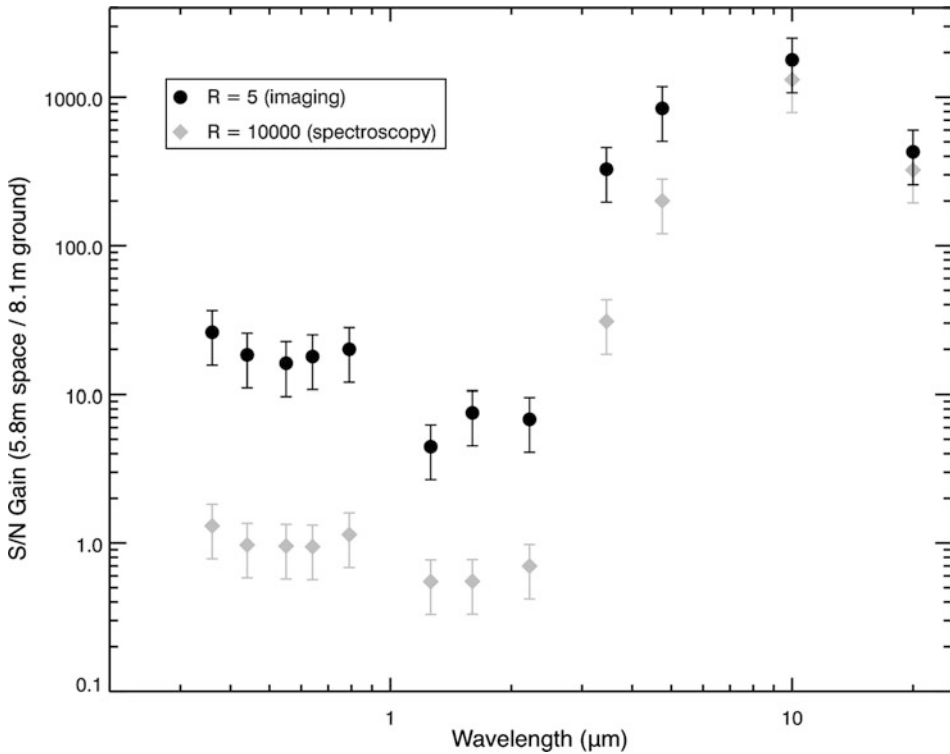
■ Fig. 9-4

Seeing and diffraction limits for different sky conditions and telescope apertures. The seeing-limited curves assume a seeing FWHM of 1 and 0.5'' at 0.55 μm . The diffraction-limited curves are shown for five optical configurations. Actual measurements of image FWHM using a variety of instruments and observatories are also shown as symbols. The vertical offset in the HST data is due to the differing measurement uncertainties of the ACS and WFC3 instruments

The individual symbols in [Fig. 9-4](#) represent actual measurements of image widths at several wavelengths using a variety of instruments and observatories. There are wavelength regimes where diffraction-limited observations are achieved from both space and ground observatories. Diffraction-limited imaging becomes extremely difficult for large ground-based telescopes below about 1.2 μm , though, due to the current limitations of AO systems. Space-based systems depart from the diffraction limit at shorter wavelengths, usually due to alignment and surface errors in the telescope optics.

1.4 A Sensitivity Comparison

A comparison of the sensitivity performance between a large ground-based telescope with a diameter of 8.1 m (similar to each of the telescopes in the Gemini Observatory) and a space-based telescope with an effective aperture diameter of 5.8 m (with equivalent collecting area to the JWST mission) is shown in [Fig. 9-5](#). Two cases are shown: imaging, where the resolving



■ Fig. 9-5

The ratio of the S/N for a space-based telescope to a ground-based observatory is shown as a function of wavelength and spectral resolution. The aperture chosen for the space-based telescope, 5.8 m, corresponds to the effective collecting area of JWST. The aperture chosen for the ground-based telescope, 8.1 m, corresponds to that of the Gemini Observatories in Hawaii and Chile. The assumptions for the calculation are listed in [Sect. 1.4.1](#). The error bars represent a $\pm 40\%$ variation in the S/N ratio

power is assumed to be 5 ($R = 5$), and high-resolution spectroscopy, where $R = 10,000$. The comparison assumes that both systems are observing the same point source and that they achieve the best delivered image quality from [Fig. 9-4](#). Both observatories are assumed to have comparable detectors and identical instrument throughputs, filter, and spectrometer band passes. The calculation is a variation on the method discussed by Gillet and Mountain (1998), and the assumptions are detailed in [Sect. 1.4.1](#).

There are two limiting observational regimes: background-limited observations and detector-limited observations. The regime is determined by the respective telescope aperture size, the background as shown in [Fig. 9-3](#), and the delivered image width as illustrated in [Fig. 9-4](#). The contributions of the respective background noise, detector noise, and source noise for three illustrative wavelengths are given in [Table 9-1](#) for the assumptions summarized in [Sect. 1.4.1](#).

For optical and NIR imaging ($R = 5$), [Table 9-1](#) shows that for equal observing times of 1,000 seconds, the observations from space and ground will be background-limited (and

■ Table 9-1

Contributions of background, detector (dark current and read noise), and source shot noise for three representative wavelengths, for hypothetical space-based and ground-based systems. Both imaging and high-resolution spectroscopy cases are shown. All noise values shown are photoelectrons per pixel for an integration time of 1,000 seconds

Imaging, R = 5			
Wavelength	0.55 μm	1.6 μm	10 μm
Source AB magnitude	30	28	24
Space system (t = 1,000 s, D = 5.8 m)			
Pixel scale (arcsec/pix)	0.022	0.035	0.217
Detector noise (e^-/pix)	3	7	12
Background shot noise (e^-/px)	36	48	2,631
Source shot noise (e^-/px)	3	8	42
Ground system (t = 1,000 s, D = 8.1 m)			
Pixel scale (arcsec/px)	0.300	0.038	0.156
Detector noise (e^-/px)	3	7	12
Background shot noise (e^-/px)	1,094	1,396	$6 \cdot 10^6$
Source shot noise (e^-/px)	17	44	197
Spectroscopy, R = 10,000			
Wavelength	0.55 μm	1.6 μm	10 μm
Source AB magnitude	24	24	22
Space system (t = 1,000 s, D = 5.8 m)			
Pixel scale (arcsec/pix)	0.022	0.035	0.217
Detector noise (e^-/pix)	3	7	12
Background shot noise (e^-/px)	1	1	29
Source shot noise (e^-/px)	1	1	3
Ground system (t = 1,000 s, D = 8.1 m)			
Pixel scale (arcsec/px)	0.300	0.038	0.156
Detector noise (e^-/px)	3	7	12
Background shot noise (e^-/px)	12	3	$7 \cdot 10^4$
Source shot noise (e^-/px)	2	2	3

■ = detector-limited; ■ = background-limited

(► 9.1) applies). In this regime, a diffraction-limited telescope in space will deliver a higher signal-to-noise observation than a comparable telescope on the ground over a wide range of wavelengths as shown in ► Fig. 9-5. At wavelengths below 1 μm , it is assumed that the ground-based telescope is seeing-limited since current AO systems cannot correct for atmospheric turbulence over a substantial field of view, as illustrated in ► Fig. 9-4. At these shorter wavelengths, ► Fig. 9-5 shows that the space-based S/N is higher by a factor of 10 or more. At wavelengths above 2 μm , the AO system produces diffraction-limited images for the ground-based system, but the background increases sharply and offsets the improved image quality, reducing the ground-based S/N gain. At a wavelength of 10 μm and beyond, the S/N of the space-based system is 1,000 times greater than can be achieved with a diffraction-limited Gemini-like system.

For spectroscopy, the background flux becomes dispersed across the detector focal plane and reduces the background seen by each pixel. The detector noise characteristics can become important, reducing the space-based advantage as [Table 9-1](#) demonstrates. For wavelengths between 1 and 2 μm with a resolving power of 10,000, the ground-based system is somewhat more sensitive than the space-based system, as shown in [Fig. 9-5](#). For wavelengths below 1 μm , the ground-based system loses its advantage because of the lower delivered image quality. Beyond 3 μm , even at resolving powers as high as $R = 10,000$, the ground-based observations become background-limited.

1.4.1 Method and Assumptions

In [Fig. 9-5](#), theoretical S/N values were calculated and compared for two hypothetical telescopes: one in space and the other on the ground. It was assumed that the two telescopes considered are able to perform observations in the optical/infrared regime. It was also assumed that they are equipped with instruments that have the same characteristics, such as quantum efficiency, readout noise, and dark current in both the optical and infrared, as listed in [Table 9-4](#).

An exposure time was estimated for a point source with a flux density that achieves a S/N of 10 in a given filter when observed using the space-based telescope. The exposure time was then used to calculate the S/N achieved by a larger ground-based telescope that observes the same source.

Twelve different spectral bands were selected to perform the analysis, covering the optical-IR wavelength range. In each band, a stellar AB magnitude was selected as shown in [Table 9-2](#) to keep the total exposure times below 10^5 s. By definition, the relationship between AB magnitude and stellar flux is given by:

$$AB(\text{mag}) = -2.5 \log(f_\nu) - 48.60, \quad (9.4)$$

where f_ν is the intrinsic source flux in units of Janskys (10^{-26} W/m² Hz) (Oke 1974).

■ Table 9-2

Assumed AB magnitudes of the sources used in the S/N comparison shown in [Fig. 9-5](#)

Central wavelength (μm)	Assumed source AB magnitude, $R = 5$	Assumed AB source magnitude, $R = 10,000$
0.36	30	24
0.44	30	24
0.55	30	24
0.64	30	24
0.79	30	24
1.26	28	24
1.6	28	24
2.22	28	24
3.45	28	24
4.75	28	24
12	24	22
20	22	16

The S/N of an observation of a stellar source calculated with an exposure time is defined as

$$S/N = \frac{I_s t}{N(t)}. \quad (9.5)$$

The number of photoelectrons per unit time from the source is given by

$$I_s = A_c \int f_\nu \cdot q_e \cdot T_a \cdot T_t \cdot T_i \cdot f \cdot \frac{1}{h\nu} \cdot d\nu, \quad (9.6)$$

where A_c represents the telescope collecting area, q_e is the detector quantum efficiency, T_a is the atmospheric transmission, T_t is the telescope transmission, T_i is the instrument transmission, f is the fraction of source photons collected (determined by the number of and size of pixels used to capture the image), h is the Planck constant, and ν is the frequency of the detected electromagnetic wave.

Expressing A_c in square meters, f_ν in Janskys, and assuming a narrow-band filter so that $\delta\nu \rightarrow \Delta\nu$ gives

$$I_s \sim 1.5 \cdot 10^7 A_c q_e T_a T_t T_i \frac{1}{R} f_\nu f \quad \text{electrons/s}, \quad (9.7)$$

where R is the spectral resolution.

The measurement noise in (9.5), $N(t)$, is a function of time and is given by

$$N(t) = \sqrt{I_s t + I_{bg} t + npix I_{dc} t + npix N_r^2} \quad \text{electrons/s}, \quad (9.8)$$

where I_{bg} represents the background flux:

$$I_{bg} = 1.5 \cdot 10^7 A_c q_e T_t T_i \frac{1}{R} \beta_\nu \Omega \quad \text{electrons/s} \quad (9.9)$$

and β_ν is the sky background surface brightness in units of Jy/arcsec², Ω is a solid angle that encloses the fraction f of the point source flux, $npix$ is the number of pixels required to cover Ω , I_{dc} is the detector dark current in units of e⁻/s/pix, and N_r is the read noise in units of e⁻/pix.

Assuming a S/N equal to 10 and a given AB stellar magnitude in a given filter, a total exposure time can be estimated for a space-based telescope from (9.5). Once the exposure time for a space-based telescope is known, the S/N of the same source as observed with a ground-based telescope can be calculated. The ratio of both S/N values is presented in Fig. 9-5.

Assumptions used for the calculations are listed in Tables 9-3 and 9-4. The sky background surface brightness (β_ν) that was used for the calculations is shown in Fig. 9-3.

It was also assumed that k observations with exposure time t can be combined with the resultant S/N given by

$$S/N(t, k) = S/N(t) \sqrt{k}. \quad (9.10)$$

1.5 Comparison of Exposure Times when Imaging Faint Objects

An important advantage of space-based telescopes over ground-based is the ability to probe to significantly fainter magnitudes for sources that are small and faint compared to the sky background, such as the high-redshift galaxies found in the Hubble Ultra-Deep Field (UDF)² shown


²<http://www.stsci.edu/hst/udf>

■ Table 9-3


List of the assumptions used in the S/N comparison presented in  Fig. 9-5



	Space-based	Ground-based
Aperture diameter	5.8 m	8.1 m
FWHM	Diffraction-limited	Seeing-limited, $\lambda < 1 \mu\text{m}$, AO-limited, $1 < \lambda < 3 \mu\text{m}$, diffraction-limited $\lambda > 1 \mu\text{m}$
Pixel scale	FWHM/2	FWHM/2
npix	4×4 for $R = 5$ 2×2 for $R = 10,000$	4×4 for $R = 5$ 2×2 for $R = 10,000$
Fractional flux (f)	0.84 for $R = 5$ 0.48 for $R = 10,000$	0.84 for $R = 5$ 0.48 for $R = 10,000$
Instrument transmission (T_i)	0.9 for $R = 5$ 0.5 for $R = 10,000$	0.9 for $R = 5$ 0.5 for $R = 10,000$
Atmospheric transmission (T_a)	1	0.6–0.95 depending on λ
Telescope transmission (T_t)	1	1

■ Table 9-4

Detector assumptions used in the S/N comparison presented in  Fig. 9-5, for both ground- and space-based observatories

Wavelength range	Quantum efficiency (q_e)	Dark current (I_{dc})	Read noise (N_r)	Assumed detector technology
0.3–1 μm	0.7	0.0008 $\text{e}^-/\text{s}\cdot\text{pix}$	3 e^-/pix	CCD
1–5 μm	0.8	0.006 $\text{e}^-/\text{s}\cdot\text{pix}$	7 e^-/pix	HgCdTe
5–20 μm	0.6	0.012 $\text{e}^-/\text{s}\cdot\text{pix}$	5 e^-/pix	Si:As

in  Fig. 9-6. This advantage allows space-based telescopes to reach very high sensitivities in reasonable exposure times, while achieving similar depths from the ground may take up to 100–1,000 times longer. The primary reasons for this are the ease with which diffraction-limited imaging is achieved from space, and the lower sky background attainable from space; these two effects contribute to different extents in the optical and near-infrared (NIR) wavelengths.

A useful metric for capturing the relative performance of different telescopes is the exposure time required to reach a given signal-to-noise ratio, as a function of the incident flux from an astronomical target of interest. An expression for S/N is given in  9.5) and  9.7), and some of the scaling laws that result are summarized here:

- For sources much brighter than the sky background (i.e., $T_a f_v f \gg \beta_v \Omega$), the exposure time required to reach a fixed S/N scales as $(T_a f_v f)^{-1}$ (for a given telescope aperture, sky background and spatial resolution, and assuming also that detector read-noise and dark current are negligible compared with the total counts from the sky and the source)
- For sources much fainter than the sky background (i.e., $T_a f_v f \ll \beta_v \Omega$), the exposure time required scales as $(T_a f_v f)^{-2}$ (again keeping all other variables as above)
- For a given source flux, the exposure time required to reach a particular signal-to-noise scales as $(S/N)^2 (\beta_v \Omega) (A_c)^{-1}$ (again under the condition that detector read-noise and dark current are negligible compared with the total counts from the sky and the source)

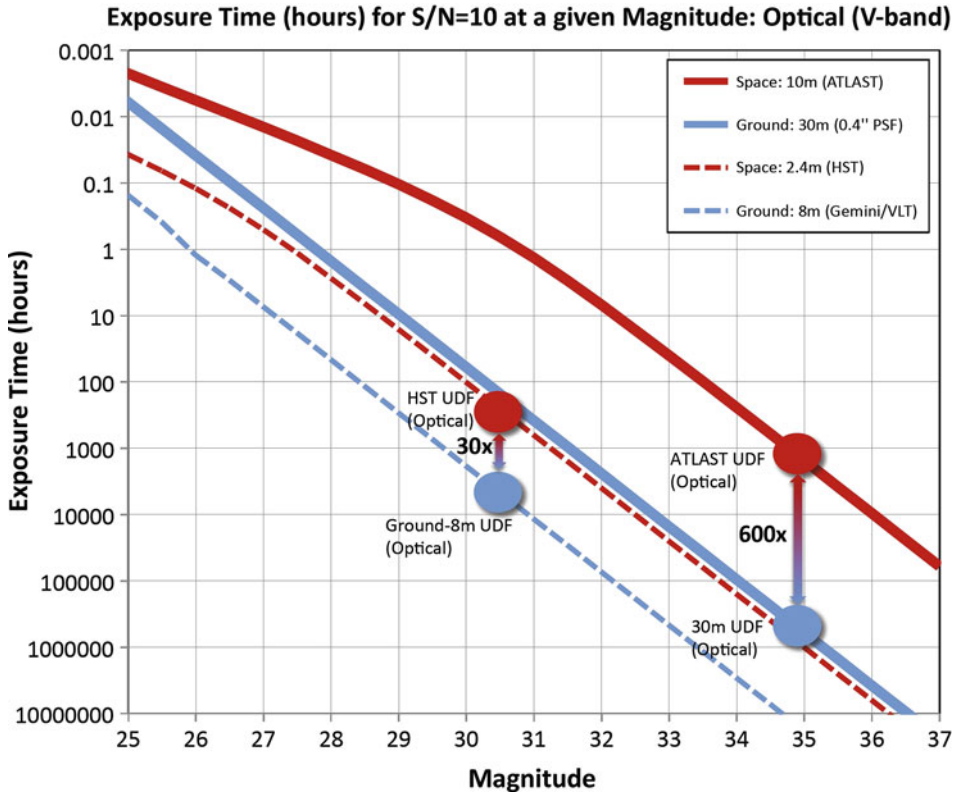


■ Fig. 9-6

The Hubble Ultra-Deep Field imaged by HST. This field is 186×186 arcsec across and was taken with a total exposure time of approximately 1,000,000 s. This image contains over 10,000 galaxies, some of which are only 0.2 arcsec across and extremely faint ($AB \sim 30.5$ magnitudes), at distances up to almost 13 billion light-years. HST, with its 2.4 m diameter mirror and $0.1''$ PSF, was able to obtain this depth with a total of about 12 days of exposure time; an 8 m ground-based telescope such as Gemini, with its much broader optical PSF, would require almost 5 years of nightly observations (including 50% efficiency due to weather) to reach the same depth, at lower resolution. Space-based telescopes provide an enormous improvement over ground-based telescopes in the regime of very small, faint sources that require the longest possible exposure times

These scaling relations can be used to examine the relative performance of ground-based versus space-based telescopes in different regimes, in particular for very small, faint sources that are a small fraction of the background flux, where much of the forefront astronomical research is focused.

At optical wavelengths, the primary advantage of space-based telescopes is the ability to reach diffraction-limited imaging over a large field of view, while a secondary advantage is the somewhat lower sky background from space at those wavelengths. Thus, for example, the $0.08''$ PSF of HST in the optical enables it to achieve depths of $AB \sim 30.5$ magnitudes in the Hubble UDF where the total investment was ~ 300 h of exposure time; by comparison an 8-m ground-based telescope such as Gemini typically achieves a $0.8''$ PSF at visible wavelengths for at least 70% of the time. Therefore, although Gemini may have a collecting area 11 times larger than that of HST, its PSF samples an area that is 100 times larger than that of HST, thereby dramatically increasing the sky background contribution within the unresolved PSF. Moreover, the sky from the ground is an additional factor of ~ 3 times brighter at visible wavelengths than that observed by HST. Therefore, as shown in [▶ Fig. 9-7](#), a Gemini-class ground-based telescope would take ~ 27 times longer (a total of 8,100 h investment, or ~ 5 years of continuous observing for 8 h per



■ Fig. 9-7

Comparison between the performance of different telescopes (HST, JWST, a hypothetical 10-m space telescope, versus 8-m and future 30-m ground-based telescopes), parameterized as the exposure time required to achieve a S/N of 10 as a function of magnitude, for optical wavelengths (V-band, or 500 nm)

night, assuming a realistic fraction of 50% usable nights) to reach comparable depths to the 300 h investment on the HST Ultra Deep Field and at a much lower spatial resolution.

Similarly, examining the performance of a hypothetical 10-m UV/optical telescope from space (ATLAST) located at L2, which would be expected to achieve a diffraction-limited PSF of $0.012''$ at blue optical wavelengths, it is found that a plausible investment of 1,000 h of time on an Ultra Deep Field could achieve limiting depths beyond AB ~ 35 th magnitude, far deeper than has ever been explored to date and opening up completely new territory. This can be compared with a hypothetical 30-m ground-based telescope, achieving a PSF of $0.4''$ at optical wavelengths: although the collecting area is nine times larger than ATLAST, the PSF subtends an angular area that is 1,100 times larger than the diffraction-limited ATLAST PSF, and in addition the ground-based sky is a factor of 5 times brighter than at L2 at visible wavelengths. Therefore, in order to reach the same depths achieved by a 1,000 h investment on ATLAST, a 30-m telescope on the ground with a $0.4''$ PSF at visible wavelengths would require ~ 600 times more exposure time (or 600,000 h), which is completely infeasible. (See [Fig. 9-7](#).)

At infrared wavelengths, ground-based telescopes are likely to be diffraction limited, though typically with some loss of PSF intensity. The principal advantage of space-based telescopes in infrared wavelengths is the fact that the sky background is dramatically fainter (a factor of $\sim 4,000$ times at H-band, or $1.6 \mu\text{m}$). Thus, for example, the near-infrared HST UDF (which achieved AB ~ 29 magnitudes with a ~ 300 h investment) would still require a factor of ~ 80 times longer to achieve on Gemini (i.e., many years of observing time). The difference is larger still when comparing Gemini to JWST, as shown in [Fig. 9-8](#). Finally, a NIR UDF obtained with a hypothetical 10-m ATLAST telescope should reach AB ~ 34 magnitudes for a $\sim 1,000$ h investment, while achieving the same depth with a ground-based 30-m telescope would require about 50 times longer integration (equivalent to several decades of observing time, using the same efficiency assumptions as before).

Therefore, it is clear that space-based telescopes provide a unique window to exceptionally faint depths that are completely unattainable from ground-based telescopes for small, compact targets, even when considering the largest possible telescopes currently being planned. In all

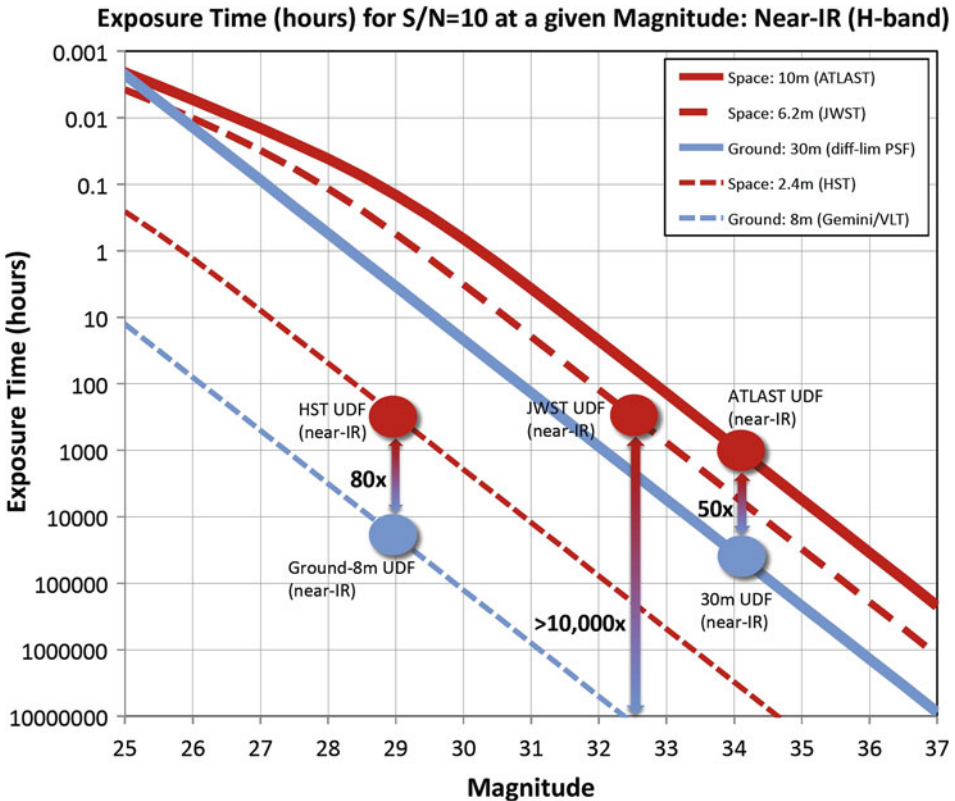


Fig. 9-8

Comparison between the performance of different telescopes (HST, JWST, a 10-m space telescope ATLAST, versus 8-m and future 30-m ground-based telescopes), parameterized as the exposure time required to achieve a S/N of 10 as a function of magnitude, for NIR wavelengths (H-band, or $1.6 \mu\text{m}$)

cases, this is primarily due to the limitations imposed by background emission, in the regime where the targets of interest are much fainter than the background: at optical wavelengths the difference in exposure times is primarily due to the much sharper resolution achievable from space (which reduces the background contribution dramatically within an unresolved aperture), while at NIR wavelengths, where ground-based telescopes will achieve diffraction-limited performance more easily, the difference is primarily due to the much lower background in space.

1.6 Very High-Contrast Imaging

One of the most exciting science goals of the “next generation” of telescopes is arguably the direct detection of habitable terrestrial planets and the search for spectroscopic biomarkers – signs of life on an exo-solar planet. In this last section on the comparisons between space-based and ground-based telescopes, the theoretical limits of a perfect groundbased telescope (corrected with AO) are examined, and the ability of the system to detect and measure an Earth-like planet in the vicinity of a star at a distance of 10 pc (~ 32 light-years) is estimated.

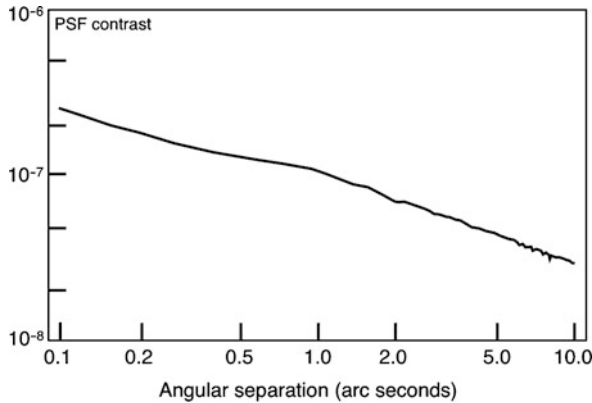
Strong motivations exist for searching for biomarkers at short wavelengths ($\lambda \sim 1 \mu\text{m}$), as discussed by Kasting et al. (2009), Lawson et al. (2009), and others. There is a strong oxygen band at 760 nm, as well as other potential biomarkers such as ozone, vegetation’s “red edge,” and Rayleigh scattering. For both the space-based and ground-based approaches, detection of such biomarkers requires the use of coronagraphic instrumentation to suppress the flux from the central star. The observations also require sufficient image quality and stability to measure the planet which is separated from its parent star by only 0.1 arcsec and has a flux of only about 10^{-10} of the star’s light.

In the laboratory, coronagraphic technologies combined with high fidelity adaptive optics to correct for imperfections in the telescope have demonstrated this level of suppression (Trauger and Traub 2007). So it becomes interesting to ask whether “near-perfect” space-based and ground-based telescopes perform comparably if fitted with a 10^{10} coronagraph.

On the ground, the necessary observations would require spatial resolutions considerably greater than the typical natural seeing (~ 0.5 arcsec), so atmospheric correction with AO would be necessary. Even with a perfect AO system, its idealized performance is fundamentally limited by the capacity to analyze the wavefront because of the finite number of photons available for wavefront sensing, which for most astronomical observations can be ameliorated with careful choice of objects, improved wavefront-sensing detectors, and/or the use of laser guide stars. However, when observations move into the regime that require measurements of 1 part in 10^{10} (required to observe earthlike planets) these fundamental limitations of AO become important (Guyon 2005).

❖ *Figure 9-9* shows the residual halo in the PSF of a perfect ground-based 30-m telescope with a perfect AO system and perfect coronagraphic instrumentation, as calculated in Mountain et al. (2009). An ideal wavefront sensor that is making theoretically optimal use of all incoming photons is also assumed. (A few existing wavefront sensor concepts do offer this level of sensitivity, but have not yet been deployed on telescopes.) Only errors due to photon noise in the wavefront sensor are considered, and the AO system is assumed to have no other source of error.

The figure shows that under nearly ideal conditions, a 30-m ground-based telescope with a perfect coronagraph and AO system will have difficulty reaching the requisite 10^{-10} contrast



■ Fig. 9-9

The residual halo in the image of a star, after suppression with a perfect coronagraph. The calculation assumes a perfect 30-m telescope observing at 760 nm and a perfect AO system with only photon shot noise in the wavefront sensor (After Mountain et al. 2009)

ratio within 0.1 arcsec of the central star because of the fundamental limits atmospheric turbulence places on such a challenging observation. The residual halo can be subtracted further (e.g., using calibration and differential methods) but even with perfect subtraction, an exposure longer than 10 days would be required to achieve a S/N of 5 (assuming $R = 70$ and a source magnitude of 4.1 in the I band). Thus, AO techniques face a fundamental hurdle to reaching the contrast of 10^{-10} necessary for imaging and low-resolution spectroscopy at wavelengths below $1 \mu\text{m}$ of an Earth-like planet at 10 pc.

In space, achieving contrasts of 10^{10} is still a daunting technical challenge, requiring either coronagraphic techniques with optical surface and alignment errors held to low levels, or free-flying, very large, high-precision starshades. These techniques, though, require surmounting technology challenges rather than overcoming fundamental limits imposed by the physics of the Earth's atmosphere.

In summary, observations from the ground and from space will continue to complement one another in the coming decades. There are certain observations that can only be done from space: deep imaging of small faint objects, low- to medium-resolution spectroscopy at the deepest magnitude in background-limited wavelength regimes, and the very high-contrast measurements needed to characterize terrestrial exoplanets and to search for biomarkers in the habitable zone around Sun-like stars.

2 Space Observatories in the UV/O/IR

The technologies used for space telescopes in ultraviolet/optical/infrared (UV/O/IR) wavelengths are constantly evolving. In infrared wavelengths, important missions include the Infrared Astronomical Satellite (IRAS), which mapped the infrared sky in the year following its launch in 1983 in wavelength ranges from 12 to $100 \mu\text{m}$. The Infrared Space Observatory (ISO) was the first general-purpose IR observatory, observed in wavelengths from 2.5 to $240 \mu\text{m}$,

■ **Table 9-5**

Basic parameters of the HST, SST, and JWST and the proposed ATLAST missions

	HST	SST	JWST	ATLAST
Wavelength range	0.12–2.4 μm	3–180 μm	0.6–28 μm	0.11–2.5 μm
Primary mirror diameter	2.4 m	0.85 m	6.6 m (maximum dimension)	10 m
Collecting area	4.24 m^2	0.49 m^2	26.3 m^2	$\sim 64 \text{m}^2$
Launch date	April 1990	August 2003	(2018, planned)	(2025–2035)
Launch vehicle	Space transportation system (Shuttle)	Delta 7920H	Ariane 5	NASA's space launch system (SLS)
Orbit	Low Earth orbit	Solar, drift-away	Sun-Earth L2	L2 or high-earth orbit

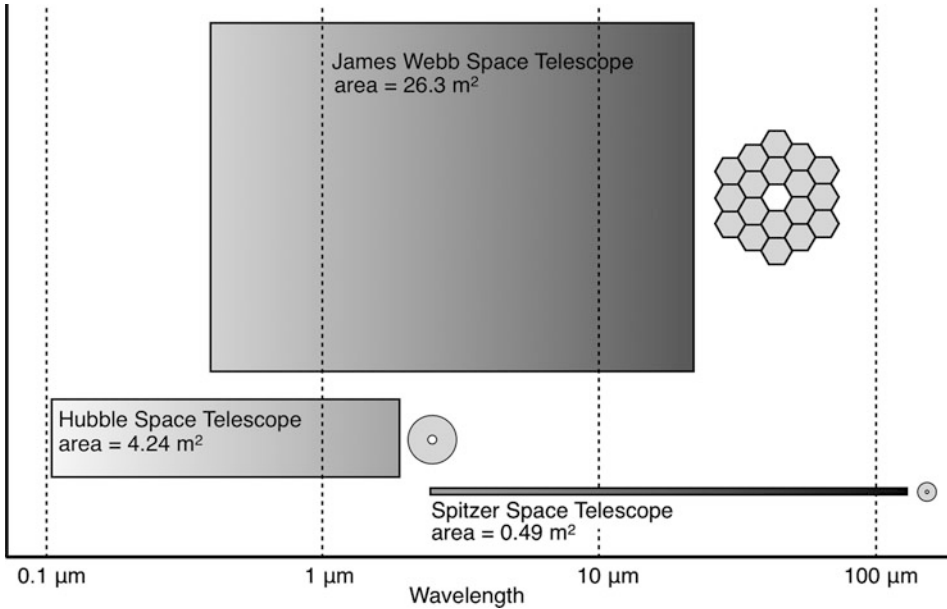
and was in operation from 1995 to 1997 (Davies 1997). The Spitzer Space Telescope (SST) is a general-purpose IR observatory launched in 2003 that used a revolutionary radiative cooling technique. It observes in wavelengths from 3 to 180 μm (Spitzer Science Center 2011). The Herschel mission, another general-purpose IR/sub-mm observatory, launched in 2009 and has an impressive primary mirror diameter of 3.5 m. It bridges the gap between IR and millimeter observations with its wavelength range of 55–672 μm (Pilbratt 2008). The James Webb Space Telescope (JWST) is currently under construction and will launch in 2018 with a maximum primary mirror dimension of 6.6 m for extremely high observing sensitivities in its wavelength range of 0.6–28 μm (Clampin 2011).

In ultraviolet and optical wavelengths, important missions include the Hubble Space Telescope (HST), a general-purpose UV/O observatory which launched in 1990 and is still providing cutting-edge science (Space Telescope Science Institute 2011). In 2003, the Galaxy Evolution Explorer (GALEX) began carrying out surveys and general-purpose observations in ultraviolet wavelengths of 0.135–0.28 μm (Oswald et al. 2004). The Kepler mission is a photometer designed to search for extrasolar planets. Launched in 2009, it is carrying out its search in optical wavelengths, 0.4–0.85 μm (Koch et al. 2010).

To illustrate the comparisons between the technologies and approaches used in space telescopes, the sections that follow give overviews of three major space telescopes: the Hubble Space Telescope, the Spitzer Space Telescope, and the James Webb Space Telescope. To illustrate how future space telescope technologies may develop, a proposed mission, the Advanced Technology Large-Aperture Space Telescope, is also discussed (Postman 2009). These particular missions were chosen because they are general-purpose observatories, and the information about each mission is available in the public domain. Primary mirror sizes, wavelength ranges, launch dates, and launch vehicles for each mission are listed in ► [Table 9-5](#). A graphical comparison of the mission's collecting areas and wavelength ranges is shown in ► [Fig. 9-10](#).

2.1 Space Observatory Elements

All space observatories have common elements. Each observatory of course has a telescope or “optical telescope assembly” (OTA) that collects photons from typically faint astronomical objects and focuses these into the science instrument package. The OTA includes the primary



■ Fig. 9-10

A representation of the relative observational parameter space occupied by JWST, HST, and SST. The area of each box is proportional to the telescope's collecting area

and secondary mirrors and their support structures. Tertiary and fine-steering mirrors also may be present. Baffles and shrouds for stray light control and any actuators that control the mirrors are also part of the OTA.

The instrument packages usually contain multiple science instruments. Each instrument typically consists of a pickoff mirror that directs light from the OTA focal plane into the instrument's optics and ultimately to its detector arrays. Instruments that are commonly included in a general-purpose observatory include imagers, spectrometers, photometers, and coronagraphs. Instrument benches, shrouds, baffles, detector, and mechanism electronics (for filter wheels and focus adjustment mechanisms) are also part of the instrument package. When there are multiple instruments, each instrument typically samples a unique portion of the OT's field of view, and the observatory is repointed to place the selected target on the correct instrument. One or more of the instruments may also provide fine guidance that increases the pointing and tracking precision of the telescope during an observation.

The OTA and instrument package are integrated into what is termed the "spacecraft bus" which includes several subsystems. Observatory pointing is coarsely measured at the spacecraft typically using star trackers. Moving between science targets is accomplished using reaction wheels, and momentum is dumped from the wheels periodically using thrusters or, if the spacecraft is in low Earth orbit and can make use of the Earth's magnetic field, magnetic momentum off-loaders. Solar panels that provide power to the observatory are also located on the spacecraft, along with the associated electronics and batteries. All command and data handling is conducted via the spacecraft; the system includes electronics, memory, and high-gain antennas

for relaying science data and commands to and from ground stations. Drive and readout electronics for mechanisms on the observatory may also be located on the spacecraft bus, which typically operates at warmer temperatures than the telescope section.

All space observatories include elaborate thermal control systems that stabilize the OTA to maintain optical quality, alignments, and background levels and to keep instruments and detectors within their operational temperature ranges. For UV and optical observatories, the OTAs and the instrument structures operate at relatively warm temperatures (typically 290 K), though the detectors are cooled to minimize detector dark current and other noise sources. The thermal control system may include multilayer insulation (MLI) blankets for protection from solar radiation, radiator panels, heat pipes, and straps to redistribute heat from electronics and mechanisms around the observatory and thermoelectric coolers for the detector arrays. Resistance heaters are often used to maintain temperature control. For infrared observatories that operate at wavelengths beyond $2\ \mu\text{m}$, it is necessary to cool the OTAs and instrument structures so that their intrinsic thermal emission does not overwhelm the faint astronomical infrared signals the observatory is trying to detect. The cooling can be accomplished passively (using carefully designed sunshields, radiators, and material choices) and/or actively (using reservoirs of cryogens, cryocoolers, or thermoelectric coolers).

2.2 Hubble Space Telescope (HST)

The Hubble Space Telescope was launched from the Space Shuttle Discovery in April of 1990 into a low Earth orbit (LEO) and is still operating at the time of this writing (2011). HST observes in the UV, optical, and very near-infrared wavelength regimes ($0.12\text{--}2.4\ \mu\text{m}$). Its primary mirror diameter is 2.4 m. A 2009 photograph of HST is shown in [Fig. 9-11](#).

HST has seen further and more sharply than any optical/UV/NIR telescope before it. Unlike some other astronomical observatories that were dedicated to a specific goal, HST has and continues to contribute to all areas of astronomical research. In its 21 years of operation to




Fig. 9-11
Photograph of the Hubble Space Telescope (NASA 2009)


date, findings enabled by HST have been published in more than 9,800 scientific papers in refereed journals, and these papers generated more than 362,000 citations. The following five achievements, for which HST observations have been absolutely crucial, deserve special note:

1. The unambiguous confirmation through the detection of very distant supernovae that the cosmic expansion is accelerating, and the most stringent constraints to date on the nature of the “dark energy” that appears to be propelling the acceleration.
2. The reduction of the uncertainty in the value of the Hubble constant to the 3% level.
3. The wealth of information about galaxy formation and evolution, as well as about the history of the cosmic star formation rate, achieved through the Hubble Deep Fields.
4. The determination of the composition of the atmospheres of extrasolar planets, and the direct visible-light imaging of such a planet.
5. The determination of the large-scale, three-dimensional distribution of dark matter through gravitational lensing.

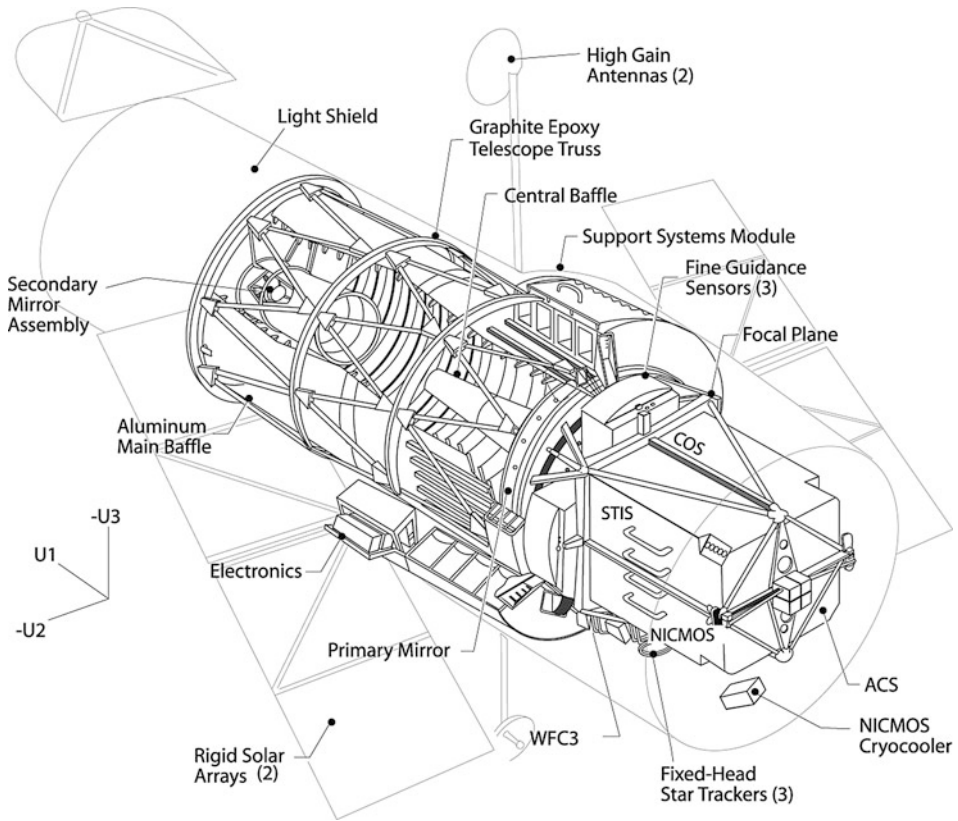
HST was placed into LEO so that astronauts could periodically service it. The five servicing missions that have been carried out since launch have given the telescope an extraordinarily long lifetime of more than 20 years. The servicing missions have made repairs to the telescope, but they have also substituted newer, better technologies for older ones, continuously updating HST’s capabilities as an observatory.

HST set several technological milestones when it was launched. It was the first general-purpose observatory in space, and the first mission that was designed to be serviceable. Its large, lightweighted primary was unique (at least for civilian missions), as was the stringent requirement on precision pointing and tracking using interferometric fine guidance sensors. Its CCDs were larger and had lower noise than had previously been flown.

A schematic of the HST observatory is shown in  Fig. 9-12. The OTA includes all the telescope components: the primary and secondary mirrors, the metering truss that sets the secondary-to-primary spacing, and the main, central, and secondary baffles. The main ring assembly holds the loads of all the OTA components and connects the OTA to the Forward Shell (FS) that encloses it. Magnetic torquers, two high-gain antennas, and the solar arrays are attached to the exterior of the FS. A light shield and aperture door are located on the front of the FS. An equipment section surrounds the main ring and contains the reaction wheel assemblies, communications system, batteries, and other electronics. The Aft Shroud (AS), located behind the main and equipment rings, contains the science instruments, the three fine guidance sensors (FGS), the star trackers, and the gyroscopes. The aft bulkhead of the AS carries a low-gain antenna and two coarse Sun sensors. The light shield, forward shell, equipment section, and aft shroud all have exterior handrails, foot restraints, and access doors to allow for human servicing (Lockheed Missiles and Space Company 1985).

As of late 2011, HST has four operational science instruments as summarized in  Table 9-6. (The Near Infrared Camera and Multi-Object Spectrograph, or NICMOS, is on board but is not currently active because its cryocooler is not operating.) Many other instruments have been installed, used, and then removed during servicing missions.³ The fine guidance sensors on HST have occasionally been used to perform scientific observations as well; they can be used to make very precise astrometric measurements, a capability that has been used to detect faint asteroids in the solar system by the shift in the background star’s centroid as the asteroids pass in front of the star.

³http://www.stsci.edu/hst/HST_overview/instruments



■ Fig. 9-12

A schematic of the Hubble Space Telescope showing its major components (Space Telescope Science Institute 2011)

■ Table 9-6

A summary of science instruments on HST that are currently in use

HST instrument	Description
Space telescope imaging spectrograph (STIS)	Spectroscopy, imaging, and coronagraphy, $\lambda = 115 \text{ nm} - 1 \mu\text{m}$
Cosmic origins spectrograph (COS)	Point-source spectroscopy, $R = 2,000 - 24,000$, $\lambda = 115 - 320 \text{ nm}$
Advanced camera for surveys (ACS)	Wide-field imaging, $\lambda = 115 \text{ nm} - 1.1 \mu\text{m}$
Wide-field camera 3 (WFC3)	Imaging, $\lambda = 200 \text{ nm} - 1.7 \mu\text{m}$

2.3 Spitzer Space Telescope (SST)

The Spitzer Space Telescope (SST) was launched in August 2003 in a Delta 7920H rocket into a drift-away solar orbit (📍 Fig. 9-13). (Prior to launch, the mission was known as the Space Infrared Telescope Facility, or SIRTF.) The telescope has a primary mirror diameter of 0.85 m



■ Fig. 9-13

A photograph of the Spitzer Space Telescope in a cleanroom (NASA 2003)

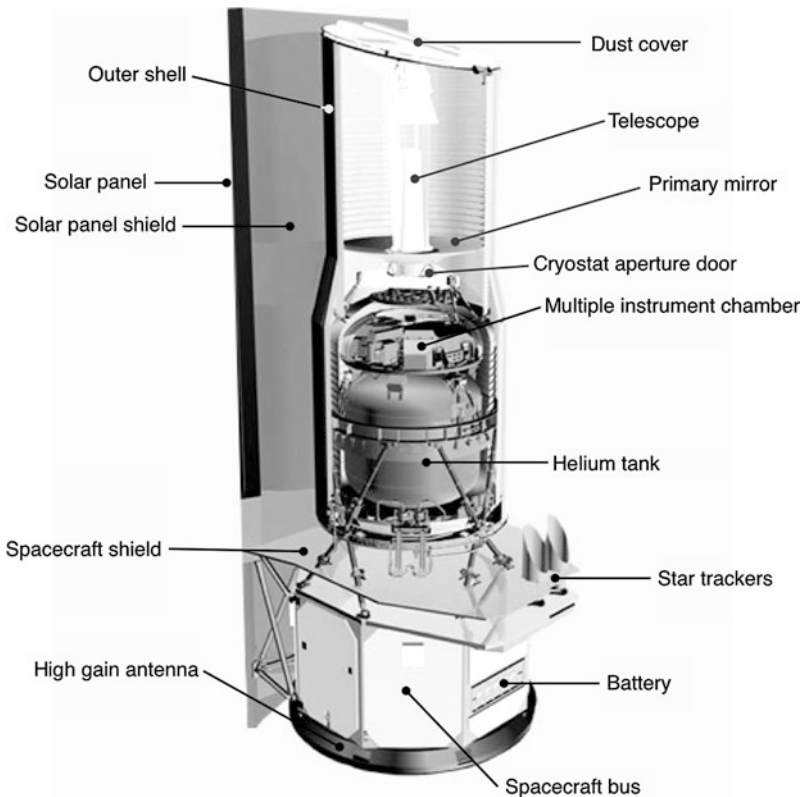
and observes in wavelengths from 3 to 180 μm . The mission carries three instruments that are summarized in [Table 9-7](#). The telescope was cooled to 4–10 K using a combination of passive and active cooling methods. The mission's superfluid helium coolant was exhausted in May of 2009. SST then moved into its “warm mission” phase; the two shortest wavelength channels of the IRAC instrument, at 3.6 and 4.5 μm , are fully operational at the time of this writing (2011).

SST was designed to study cold and dusty environments, as well as the distant universe. More than 3,000 scientific papers based on SST results have been published in its 8 years of operation, and those papers have generated over 95,000 citations. SST has characterized the atmospheric structure and the composition of extrasolar planets. It has demonstrated that the process of planet formation is well underway within a few million years of the formation of Sun-like stars and that organic materials have been abundant in the universe for the past 10

■ Table 9-7

A brief description of the three instruments aboard the Spitzer Space Telescope

SST instrument	Description
Infrared array camera (IRAC)	Wide-field imaging, $\lambda = 3.6\text{--}8\ \mu\text{m}$
Infrared spectrograph (IRS)	Low-resolution spectroscopy, $\lambda = 5.2\text{--}38\ \mu\text{m}$ High-resolution spectroscopy, $\lambda = 9.9\text{--}37.2\ \mu\text{m}$
Multiband imaging photometer for SIRTf (MIPS)	Photometry and imaging, $\lambda = 24, 70, 160\ \mu\text{m}$ Low-resolution spectroscopy, $\lambda = 50\text{--}90\ \mu\text{m}$



■ Fig. 9-14

Schematic of the Spitzer Space Telescope (After Spitzer Science Center 2011)

billion years. It has also determined the masses and ages of galaxies when the universe was just 1 billion years old. While in its current “warm mission” phase, SST is engaged in an effort to further reduce the uncertainty in the value of the Hubble constant.

A schematic of the SST observatory is given in ► Fig. 9-14. The Cryogenic Telescope Assembly (CTA) includes the telescope, the superfluid helium cryostat, the outer shell, and the Multiple Instrument Chamber (MIC) which is mounted to the helium tank. The MIC contains the three science instruments and a pointing control instrument (PCRS). The spacecraft includes the instrument electronics, solar panel, high-gain antenna for communications, star

trackers, four reaction wheels, and nitrogen propellant that is used for off-loading momentum that has built up in the reaction wheels.

The SST was the first to use a thermal design that passively cooled the telescope portion of the mission. The design drastically reduced the amount of coolant required and significantly lowered the total mass and volume of the mission. SST was also the first to use a lightweight beryllium primary mirror. In addition, SST's instruments used infrared detector arrays with higher quality and larger formats than previous infrared missions (Gehrz et al. 2007).

2.4 James Webb Space Telescope (JWST)

The James Webb Space Telescope (JWST), currently scheduled for a 2018 launch, will provide imaging and spectroscopy in wavelengths from 0.6 to 28 μm . It will launch in an Ariane 5 rocket and will orbit about L2. The segmented primary mirror is 6.6 m in diameter point-to-point and will provide a collecting area of 26.3 m^2 (Lightsey et al. 2012). A schematic of the system is shown in [Fig. 9-15](#).

The James Webb Space Telescope (JWST) will be the natural successor to the iconic Hubble Space Telescope. It is expected to extend our view of the universe to regions beyond Hubble's reach. A comparison of an HST image to a simulated JWST image is shown in [Fig. 9-16](#). With its 6.6 m maximum aperture diameter and wavelength range of 0.6–28 μm , JWST will be ideally suited for observing the very first galaxies that have formed in the universe. These galaxies are believed to have contained about a million solar masses in stars. The ultraviolet light from these stars is redshifted into the NIR due to the intervening cosmic expansion. Since the first generation of galaxies has probably played a crucial role in the cosmic reionization, JWST will provide unique insights into this universal phase transition. While JWST will not be

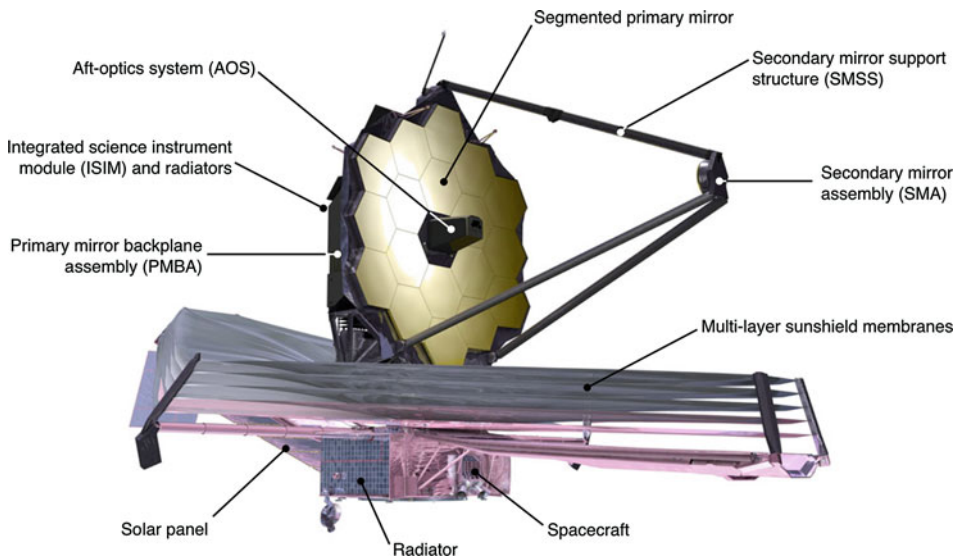
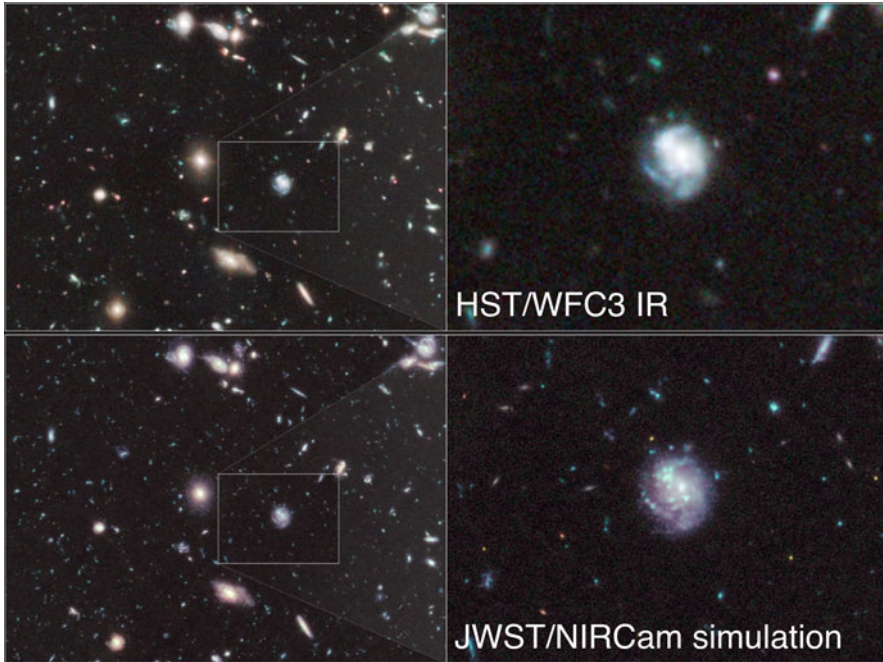


Fig. 9-15

An artist's conception of JWST (NASA 2009, labels added)



■ Fig. 9-16

JWST will obtain spectacular images with unprecedented power for discovery. The top panels show an image of the Hubble Ultra-Deep Field in the near infrared, recently taken by Hubble's WFC3 instrument. This 0.7×1 arcmin image of the universe is the deepest ever obtained. The bottom panels show a simulated image of the same region by the NIRCam instrument on JWST. JWST's higher spatial resolution improves the clarity of the image and better reveals the morphology of the sources. JWST's exquisite sensitivity – much better than that of HST – brings out faint galaxies hidden in the noise of the Hubble image (Image courtesy of Space Telescope Science Institute: M. Stiavelli, J. Kalirai, Z. Levay)

able to discern the individual first stars (Population III stars), it may be able to see the supernova explosions (Pair-Instability Supernovae) of these stars.

Going from the very large, cosmological scale, to the small, planetary scale, JWST will be able to identify extrasolar planets in habitable zones that have liquid water on their surfaces. This will be a giant step toward finding the necessary ingredients for life on extrasolar planets. Clearly, JWST will address a rich host of phenomena at intermediate scales, as well. These will range from the assembly and evolution of galaxies to the birth of stars and the formation of protoplanetary systems. There is little doubt that JWST will provide new and valuable insights into the origin of the cosmos and our place within it.

The main components of JWST are sketched in ► Fig. 9-15. The OTA includes the 18 primary mirror segments, the primary mirror support structure, the secondary mirror, the secondary mirror support structure, and the Aft Optical System (AOS). The AOS contains the tertiary mirror, the fine-steering mirror (FSM), and baffles and apertures for stray light control. The Integrated Science Instrument Module (ISIM) carries the observatory instruments and is

located behind the primary assembly. ISIM is enclosed by radiators that cool the instruments to about 40 K. A deployable tower connects the telescope to the spacecraft, which carries the solar array, reaction wheels, star trackers, warm electronics, thrusters, and a high-gain communications antenna. A multilayer sunshield protects the observatory from stray light and thermal loads from the Earth and Sun.

JWST's design builds on the extensive experience from HST, SST, and to some extent, from the Chandra X-ray telescope. It will demonstrate the feasibility of segmented primary mirrors that can fold to fit into available launch vehicles. The deployable multilayer sunshield is unique to space observatories, though commercial and military radar and communication satellites have used deployable structures for many years. JWST will be the first mission in which full on-orbit alignment of the optics is carried out. Finally, JWST will be the first observatory to demonstrate a thermal design that relies almost entirely on passive cooling methods, building on SST's success with passive cooling techniques. JWST will carry four science instruments, as listed in [Table 9-8](#); of these, only the MIRI instrument is actively cooled.

2.5 Advanced Technology Large-Aperture Space Telescope (ATLAST) and Other Future Missions

In looking to the future, it is important to understand what the intrinsic science drivers might be for future space telescopes beyond those for the telescopes above. One compelling encapsulation of the key science themes for the twenty-first century was given by the Nobel Laureate Riccardo Giacconi, "to relate causally the physical conditions during the Big Bang to the development of RNA and DNA" which can be distilled into key questions such as:

- How did the universe originate and what is it made of?
- How unique was our occurrence? Are we alone?

Using today's almost perfect detectors ($QE \sim 1$) and exploiting the low background available in space, the term QE_λ/B_λ in [\(9.1\)](#) is negligible. Combining [\(9.1\)](#) and [\(9.3\)](#) shows that for diffraction-limited observations from space, signal-to-noise ratio is proportional to the square of the aperture diameter:

$$S/N \propto \frac{D^2}{\lambda}. \quad (9.11)$$

Table 9-8

Instruments planned for the James Webb space telescope

JWST instrument	Description
Near-infrared camera (NIRCam)	Wide-field imager, $\lambda = 0.6\text{--}5 \mu\text{m}$
Mid-infrared instrument (MIRI)	Imaging, low- and medium-resolution spectroscopy, coronagraphy, $\lambda = 5\text{--}28 \mu\text{m}$
Near-infrared spectrometer (NIRSpec)	Multi-object spectrograph, $\lambda = 0.6\text{--}5 \mu\text{m}$
Near-infrared imaging slitless spectrometer (NIRISS)	Imager and slitless spectrometer, $\lambda = 0.6\text{--}4.8 \mu\text{m}$

Since exposure time is proportional to the square of S/N, the exposure time needed to reach a given signal-to-noise ratio decreases by D^4 . If it were possible to enable even larger telescopes than JWST, future UV/O/IR space telescopes could achieve unprecedented sensitivities and spatial resolutions (e.g., see ● *Figs. 9-7* and ● *9-8*). One such mission is the Advanced Technology Large-Aperture Space Telescope (ATLAST), which would operate in the wavelength range of 0.11–2.5 μm with segmented primary mirror diameters as large as 16.8 m. To explore lessons from previous space telescopes and the interplay in the design and operation of future space telescopes, the 10-m ATLAST concept will be considered a reference mission for this chapter.

ATLAST could be the first telescope to detect biosignatures – the spectral features generated by life – on Earth-like planets. Biosignatures on Earth are combinations of the spectral features from molecular oxygen, water, methane, and ozone. Some spatial information about the planet's surface could also be recovered by measuring time variations in the planet's brightness. In addition, the mission could carry out UV surveys of the local intergalactic medium to reveal the physics behind galaxy formation and evolution, determine star formation histories for distant galaxies, and study dark matter by measuring the proper motions of stars in galaxies (Postman 2009).

Other mission concepts with the goal of taking images and spectra of earthlike planets have been proposed. The New Worlds Observer proposes a 4-m class telescope used in combination with an external starshade operating in a wavelength range of 0.1–1.1 μm (Cash et al. 2009). Similarly, the New Worlds Probe proposes the use of an external starshade for improved high-contrast coronagraphic imaging with JWST. The mission would rely on the NIRC*am* and NIRS*pec* instruments and would operate in wavelengths longer than 0.7 μm (Lo et al. 2010). Another mission that proposes an external occulter is the Telescope for Habitable Exoplanets and Interstellar/Intergalactic Astronomy (THEIA). THEIA includes a 4-m class telescope in the wavelength range of 0.1–1 μm (Kasdin et al. 2009).

Other proposed infrared observatories include the Single-Aperture Far-Infrared Observatory (SAFIR) mission, which would study the formation of the first stars in the universe, the formation of planetary systems, and the complex molecules that might presage life in planet-forming regions (Lester et al. 2004). SAFIR would operate in the wavelength range between JWST and ground-based microwave observations (20 μm –1 mm). With a primary mirror diameter of 5–10 m and an operational temperature of 5 K, the mission would have significantly greater far-infrared sensitivity than the SST or Herschel missions. A proposed mission known as the Space Infrared Telescope for Cosmology and Astrophysics (SPICA) would carry a 3.5-m telescope and would be optimized to operate in the mid- and far-infrared (3.5–210 μm). The mission would observe in the wavelength gap between JWST and Herschel, with higher sensitivity than that of the SST (Nakagawa 2008).

3 Science Goals, Cost, and Productivity

As has been shown in the preceding sections, there is a broad observational parameter space where space telescopes are extremely competitive. In certain wavelength regimes, space-based observatories are essential. The central precept behind designing (and eventually flying) a space telescope is to ensure that the science a specific telescope can deliver is sufficiently compelling that it justifies the additional cost and complexity of the observatory.

Ideally, the design of a mission would be driven by the science objectives. In reality, though, cost is a major constraint, and the science objectives must be maximized within the available technology and, crucially, the funding limits. If the estimated cost is found to be too high in the early design phase of a mission, science objectives might be reduced and the mission redesigned accordingly. HST, SST, and JWST have been through redesigns that illustrate the types of trades made during the design phase of a space telescope. Parameters that influence a mission's scientific return and its cost (usually inversely related) include primary mirror size, launch vehicle, mass, launch volume, choice of orbit, thermal control strategy, and operation approach.

A key approach to finding the optimum combination of science return, engineering maturity versus technology development and ultimately overall mission cost is the use of “science systems engineering.” The phrase was originally coined by Nobel laureate Riccardo Giacconi:

I should define here what was meant by science systems engineering, and expression we at STScI came to use with increasing frequency. It signifies the analysis of a scientific research problem in all its dimensions, even before developing the instrumentation. Starting with a clear definition of the problem, we would design instruments capable of obtaining the necessary data, then plan how these data would be analyzed, determine what errors might occur because of the intrinsic limitations of the instruments, and define the expected statistical weights of the observations. Only when it was clear that crucial results could be obtained would we proceed with the project. The principles of science systems engineering would be applied throughout the lifetime of the project to ensure that no changes occurred that would jeopardize its scientific success. – Riccardo Giacconi, *Secrets of the Hoary Deep*

Science systems engineering is used to describe the process of designing, deploying, and operating complex space telescopes and instruments in a way that ensures that they meet their ultimate scientific objectives. The science goals and the flow-down to science requirements must be part of the trade studies and the decision processes that go on continuously throughout a mission's life. Science systems engineering is still used in the operation of the Hubble Space Telescope to ensure that it is fulfilling its designated mission (Giacconi 2008). Some of the factors that have to be considered in a science system engineering approach are discussed in the next sections.

3.1 Early Design Trades

The Hubble Space Telescope was originally conceived by Lyman Spitzer in 1946 in a famous paper for the Rand Corporation, “*Astronomical Advantages of an Extra-Terrestrial Observatory*.” Originally called the large space telescope (LST), HST was conceived of as a “120-in.” (3-m) telescope, which was large by the standards of the day. At launch, HST's primary mirror was 2.4 m in diameter. To quote from NASA's history

In the mid-1960s, NASA and its contractors conducted phased studies into the feasibility of a large space telescope. Although there was initial dissent within NASA over whether the agency should work its way up to a large-scale observatory or take one giant leap to the final product, the decision to develop the Space Shuttle program greatly improved the flexibility NASA would have in designing a space telescope. In 1971,

George Low, NASA's Acting Administrator, gave approval to the Large Space Telescope Science Steering Group to conduct feasibility studies...

Unfortunately for the program, the large space telescope's total cost was roughly estimated at \$400 to \$500 million [dollars in 1970], making it a tough sell... A mirror reduction from 3 to 2.4 meters helped bring the project down to about \$200 million, approximately half the originally expected price tag. The proposal was accepted by Congress, which granted the Large Space Telescope program funding in 1977.⁴

The Spitzer Space Telescope went through major redesigns twice during its design phase. Originally, the mission weighed 5,700 kg and was to be placed into a high Earth orbit by a Titan IV launch vehicle at a projected cost of approximately \$2.2B in 1990 dollars. The mission was radically redesigned and launched in a smaller Delta rocket into a heliocentric Earth-trailing orbit, with a payload mass of 860 kg and a cost of approximately \$0.7B in 1990 dollars. The redesign was accomplished without reducing the size of the telescope or the lifetime of the mission. The science goals of the mission were reduced somewhat and the instruments simplified, but the core science capabilities of the mission were preserved (Spitzer Science Center 2011).

JWST also went through a number of descopes in its early design phase. The original concept was an optical/infrared telescope with a 36-segment primary mirror 8 m in diameter, as sketched in [Fig. 9-17](#). In 2005, the primary mirror diameter was reduced to 6.6 m (maximum dimension) and a primary design with 18 segments was selected. The minimum wavelength for diffraction-limited imaging was raised to 1 μm . In addition to those changes, two instruments were transferred to ESA and CSA, the instruments were simplified, and detector procurements were consolidated. Later, the diffraction-limited wavelength was relaxed from the visible to the near-IR to simplify fabrication of the telescope, and one of the Tunable Filter instruments (now NIRISS) was removed to improve the mass and power margins for the mission. Optimization of the design of major space observatories goes beyond just the resizing of the primary mirror.

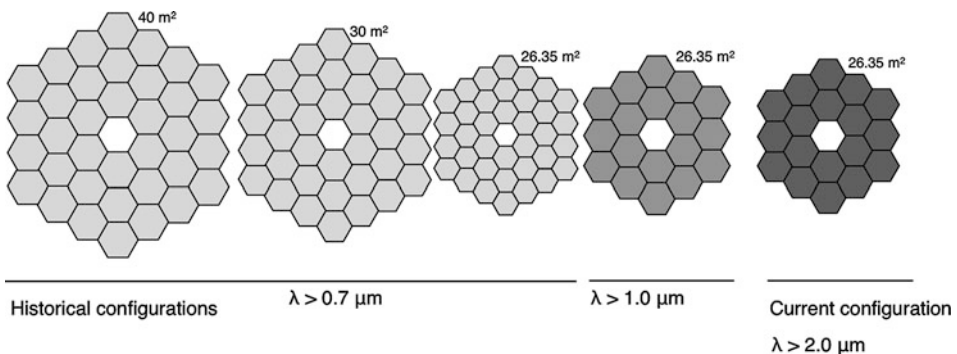


Fig. 9-17

The evolution of the JWST primary mirror configuration from a 36-segment mirror with an 8-m diameter, to its current configuration of 18 segments with a maximum dimension of 6.6 m. Configurations are proportional to one another

⁴<http://history.nasa.gov/hubble/index.html>

3.2 Mission Cost

The factors that drive a mission's cost are not perfectly known, but various models, based on parameterized descriptions of past missions, can be used to try to predict costs (Stahl 2010). In the past, the total mass of a mission has been the traditional metric to judge the cost of a space mission: putting more mass into orbit normally cost more. However, mission complexity for space-based observatories has been rising as larger systems must be fit into smaller launch volumes. A heavier, simpler mission may be less expensive than a lighter but more complex one. A cost model that attempts to account for complexity is the NASA Advanced Mission Cost Model (AMCM), which here has been adapted to report cost in 2010 dollars:

$$C = (\$2.636 \cdot 10^9) \cdot (\text{mass}/10,000 \text{ kg})^{0.654} \cdot (1.555^{\text{difficulty level}}) \cdot (N^{-0.406}), \quad (9.12)$$

where C is the cost to launch of the mission in 2010 dollars, N is the number of identical flight systems manufactured, which has always been 1 for space-based astronomical observatories, and the difficulty level ranges from -2 for “very low” to 2 for “very high.”

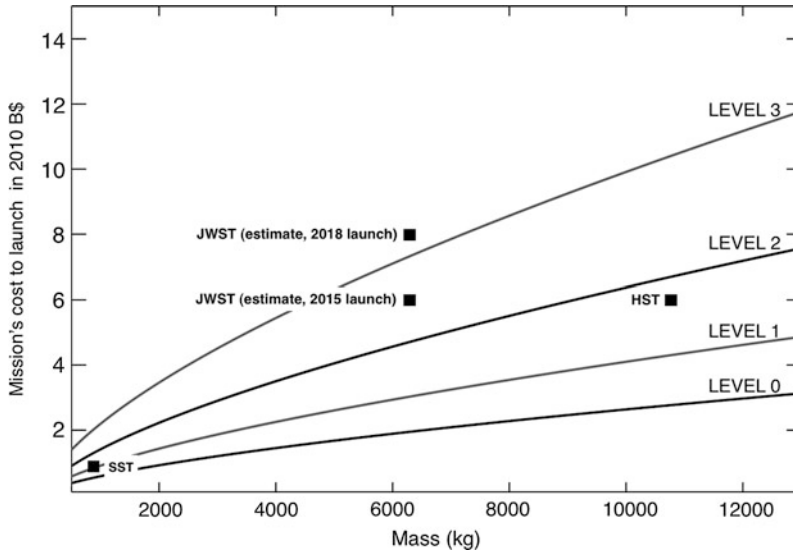
Figure 9-18 shows the resulting relationship between cost, mass, and difficulty level. The curve for a difficulty level of 3 is shown in an ad hoc attempt to account for the increased complexity of systems such as JWST with deployable segmented primary mirrors, large deployable multilayer sunshields, and the required new approach to integration and test (Casani et al. 2010a, b). No such missions were included in the historical missions from which the AMCM was created. HST has a difficulty level of nearly 2, and JWST is significantly more complex than HST.

The cost to develop, build, and launch HST is estimated at roughly \$6B in 2010 dollars, as shown in Figure 9-18. When the costs of the five servicing missions and the operating costs from 1990 through to 2014 are included, NASA has estimated that the total mission cost of HST is \$19B in 2010 dollars. SST cost roughly \$0.9B to launch and functioned as a cold infrared observatory for nearly 6 years. JWST is currently projected to cost \$8B (in 2010 dollars) from design through its launch in 2018. Although more expensive than ground-based observatories, the scientific productivity of general-purpose space-based observatories like HST and SST has been extremely high.

3.3 Mission Productivity Metrics

In the “science system engineering” paradigm, the potential scientific impact of a space telescope is a key part of the trade space when optimizing a space mission. Larger, more expensive missions like HST, SST, and JWST, are designed to have a significant “discovery” potential owing to their broad astrophysics objectives. A science mission's productivity be roughly measured by the number of peer-reviewed scientific papers produced. The number of scientific papers published each year for various observatories is shown in Figure 9-19. HST and SST have produced large numbers of papers in recent years.

Another productivity metric is the number of citations, which correlates to the value to other researchers of the papers that have been published. Missions with a narrow focus may not generate a large number of papers but might generate large citation counts because they made critical contributions to key areas of science (the Wilkinson Microwave Anisotropy Probe, or WMAP, falls into this category). Attempts are also made to measure productivity by counting the number of significant discoveries made by an observatory. The “Davidson metric”

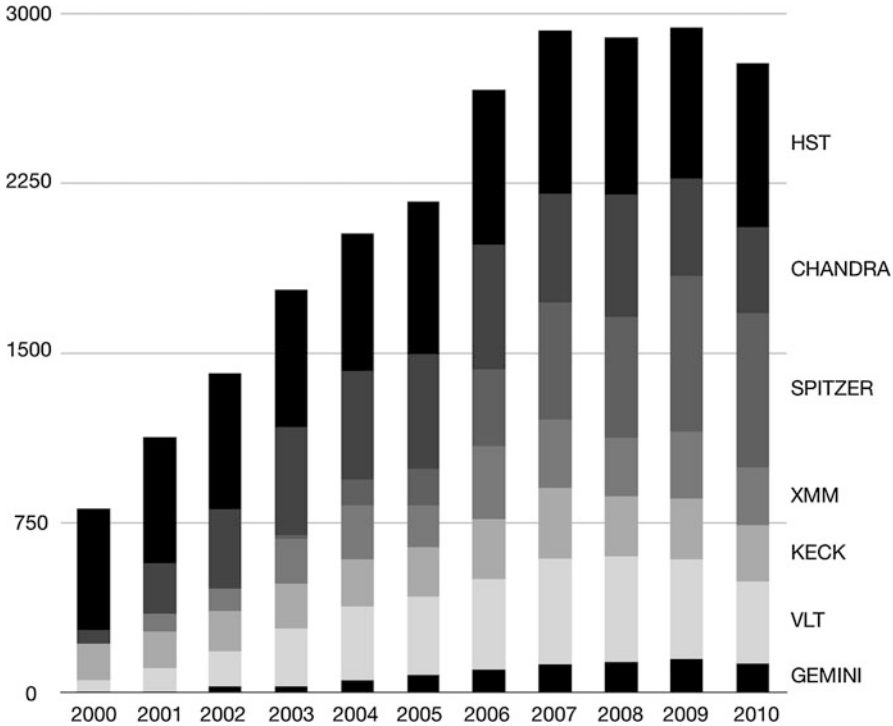


■ Fig. 9-18

The cost of a mission from birth through launch, as a function of mass and difficulty according to NASA's Advanced Mission Cost Model. The actual cost to launch values for SST and HST are shown in 2010 dollars, as are two estimates for JWST's cost-to-launch: one is a 2010 estimate by the Independent Comprehensive Review Panel (ICRP) of the minimum cost to launch JWST (Casani et al. 2010a, b), and the other is the amount currently budgeted by the JWST program as of 2011, which includes the cost of stretching out the launch date by three additional years as compared to the ICRP estimate. All values shown are 2010 dollars (Note that the AMCM was not designed to accommodate a difficulty level of 3; no missions with deployable primary mirrors, large deployable multilayer sunshields, and new approaches to I&T (Casani et al. 2010a, b) for missions of JWST's scale were included in the historical missions used to design the metric)

counts the number of stories in *Science News* that feature a significant discovery by a given mission, as shown in ► Fig. 9-20. Because *Science News* is a popular publication, the metric is also related to a mission's impact on the public (Christian and Davidson 2006).

As can be seen in ► Figs. 9-19 and ► 9-20, as the preeminent UV/optical observatory that has been in service for more than 20 years, HST has maintained a stunning level of scientific return and has had a significant impact on the public. This longevity is largely due to the five successful servicing missions that have continually repaired and upgraded the telescope and its instruments. SST has also become the preeminent IR observatory in its 6 years of cold observations, achieving high levels of scientific return despite its limited aperture size by exploiting a wavelength and sensitivity range that was inaccessible to ground-based systems and unexplored by previous space telescopes. JWST was explicitly designed to build on the scientific capabilities and legacies of both HST and SST. The science systems engineering approach was used to ensure that JWST has comparably high scientific productivity and impact, based on the choice of operational wavelength range enabled by its low operational temperature, the requirement for extremely high sensitivities and spatial resolutions enabled by its aperture size, its comprehensive suite of imaging and spectroscopic instruments, and through its operational lifetime goal of 10 years.



■ Fig. 9-19

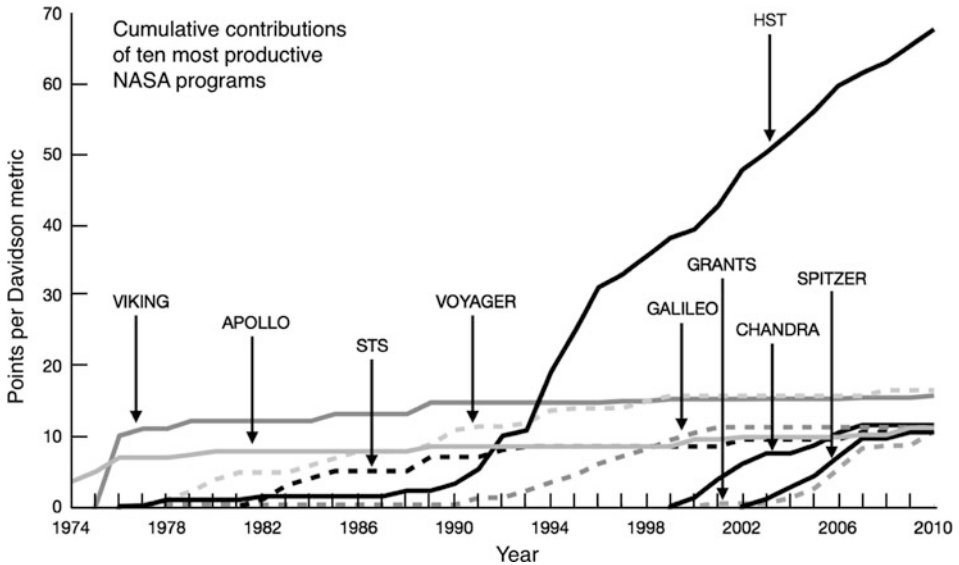
Number of scientific papers published each year for various observatories. Journals considered include: *Astrophysical Journal*, *Astronomy and Astrophysics*, *Astronomical Journal and Supplement*, *MNRAS*, *Icarus*, *Publications of the Astronomical Society of the Pacific*, *Nature*, *JGRA*, *Science*, and others (Lagerstrom 2011, private communication; Apai et al. 2010)

4 Orbits

The choice of orbit is a significant consideration, affecting the choice of launch vehicle, the background radiation levels, observing efficiency, ease of communication, and the mission lifetime. These factors, in turn, contribute to the overall mission cost. HST is in a low Earth orbit, SST is in a solar drift-away orbit, and JWST will orbit the second Lagrange point of the Earth-Sun system (L2). Other common satellite orbits include Sun-synchronous, geostationary, and high Earth orbits and are discussed by Bely (2003) and Davies (1997). (Also see Prussing and Conaway 1993; Bate et al. 1971.)

4.1 Low Earth Orbit

HST is in a low-inclination, circular, LEO at about 600 km in altitude with an orbital period of 96 min. The inclination of the orbit is about 28° , and was determined by the latitude of the launch site at Cape Kennedy, Florida. The orbital nodes precess westward by about 6° per day.



■ Fig. 9-20

Productivity of ten NASA missions according to the Davidson metric. The number of points shows the cumulative number of discoveries between 1973 and 2010 (Carol Christian 2011, private communication)

LEO has two major advantages: it is cheaper to reach than most other orbits and offers the opportunity for human servicing. Five servicing missions went to HST to make repairs and install new instruments, enabling the observatory to be scientifically productive for more than 20 years. For missions that are not serviced, LEO is still advantageous because for a fixed payload size, a smaller launch vehicle can be used to reach orbit. For a fixed cost, a larger payload can be placed in LEO than in most other orbits. LEO also allows for high-bandwidth communications to be carried out using relatively small onboard high-gain antennas.

The observing efficiency in LEO is limited by Earth occultations. Objects near the orbital plane of HST are routinely occulted for about 52 min so that HST's maximum long-term observing efficiency can never be much higher than 50% (Doxsey 2006). Objects near HST's orbital poles can be viewed continuously for several days, however. This "continuous viewing zone" was used to perform several of the HST deep-field campaigns.

Communication with a satellite in LEO must be carefully planned since the satellite can contact an individual ground station for only a few minutes at a time. HST communicates using NASA's Tracking and Data Relay Satellite System (TDRSS) (Davies 1997). The system relies on nine satellites located in geosynchronous orbits so that their locations with respect to the Earth's surface remain fixed. The antennas on each satellite are designed so that signals can be received from a wide range of angles. A ground station located in New Mexico communicates with the network. Data from HST can be transmitted between individual TDRSS satellites until it reaches a satellite that is in contact with the ground station. Commands to HST travel the reverse route.

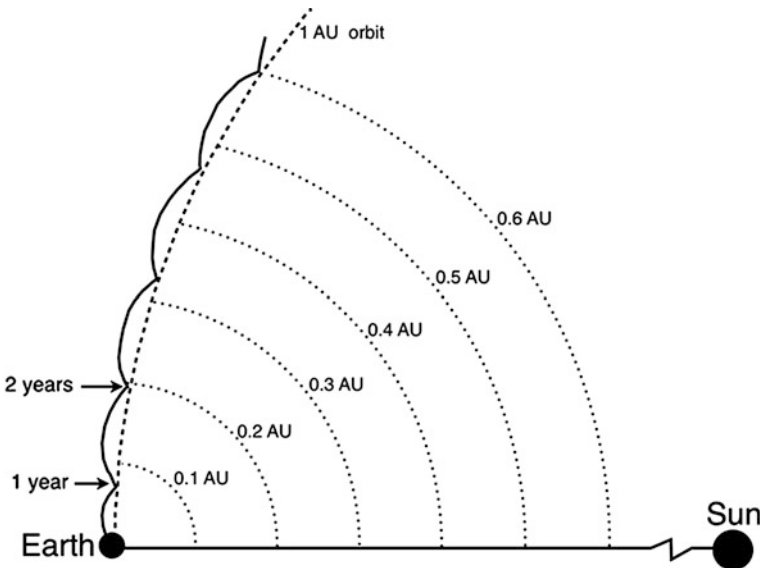
LEO's lie between about 300 and 1,000 km in altitude. Below 300 km, atmospheric drag shortens a mission lifetime to impractical levels. Above 1,000 km, there is a dense population

of high-energy particles trapped in the Van Allen belts. The inner belt contains high-energy protons and electrons (roughly 30–50 MeV) that are spiraling around in the Earth’s magnetic field. Such particles can cause false detections that obscure the desired science signals and can occasionally cause unexpected behaviors in the flight electronics. Even within the LEO range, there are some regions known to be plagued by higher densities of charged particles. The most significant region is known as the South Atlantic Anomaly (SAA) and HST is programmed to cease observations while it passes through the SAA.

HST’s low Earth orbit causes some degree of thermal instability. As it passes in and out of the Earth’s shadow and thermal load, HST is exposed to temperatures from -80 C to $+50\text{ C}$. MLI blanketing prevents the telescope structure from experiencing the full range of these temperatures, but the thermal changes that are transmitted to the telescope structure can cause small transient changes in image quality and pointing. The errors are handled with a combination of real-time and post-facto corrections.

4.2 Drift-Away Orbit

The Spitzer Space Telescope is in a solar orbit that trails the Earth, drifting away at a rate of about 0.12 AU per year. The orbit is plotted in [Fig. 9-21](#). Launch into a drift-away orbit is efficient,



■ Fig. 9-21

SST’s solar orbit as viewed in the rotating frame; the Earth-Sun line is fixed and the coordinate system is centered on the Earth. When looking back from Earth at a trailing object in an identical elliptical orbit, the object will appear to move in an elliptical pattern over the course of 1 year. SST’s orbit is slower and has a slightly different elliptical shape than that of the Earth; when viewing SST from the Earth, then, it traces a more complicated shape. Much of SST’s time is spent in the “turnaround” points sketched in the figure (After Gallagher et al. 2003)

since it requires reaching just a little more than Earth's escape velocity. No additional fuel for circularizing the orbit is required, which saves a great deal of launch mass, and no onboard propulsion is needed to maintain the satellite's position.

Because the observatory is far from the Earth, it simplifies observation planning and leads to high observing efficiencies. The orbit also removes the Earth as a heat source for the observatory, which is critical in the infrared where the thermal background from all sources must be kept low. An observatory in drift-away orbit experiences relatively constant temperatures since the Sun is always in the same relative position to the spacecraft.

Large ground antennas are required for communication after the observatory has drifted a long distance from the Earth (Gallagher et al. 2003). SST communicates using NASA's Deep Space Network (DSN), which consists of ground-based radio antennas placed around the globe to ensure that constant contact with spacecraft can be maintained despite the Earth's rotation. Each station contains large, steerable radio dishes ranging in size from 26 to 70 m. By January of 2014, SST will be so far from Earth that consistent communication with its low-gain antenna will no longer be possible, and it would be difficult to restore telescope operations if the observatory went into its safe mode.

4.3 Libration Point Orbit at Sun-Earth L2

JWST will orbit about the second Lagrange point of the Earth-Sun system as sketched in [Fig. 9-22](#). The Lagrange points are solutions to the three-body gravitational problem in the special case that all three bodies lie in a plane, the second body is in orbit about the first, and the third body has a mass that is negligible compared to the other two bodies. A small mass inserted at any of the Sun-Earth Lagrange points will stay in a nearly constant position with respect to the Earth and Sun. The L1 and L2 points are semi-stable; if the satellite were nudged forward or backward along its orbital path, the gravitational forces from the Earth and Sun would cause it to return to its original position. A nudge perpendicular to the orbital path would not be self-correcting. (The L4 and L5 Lagrange points, about 1 AU from the Earth, exist

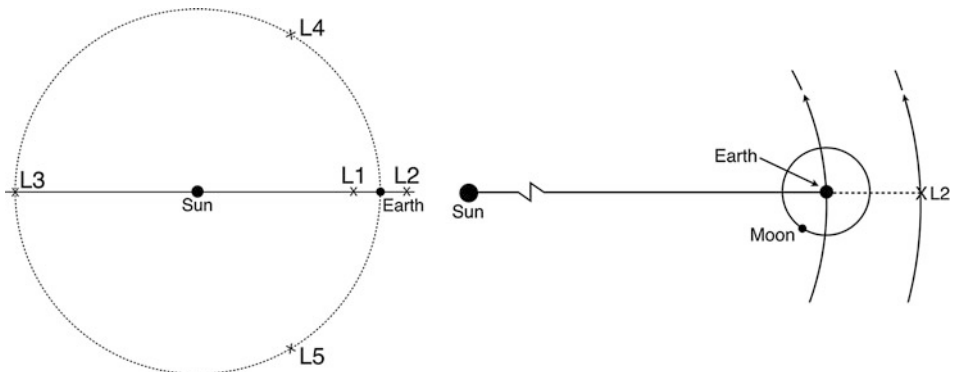


Fig. 9-22

A schematic of the Lagrange points in the Earth-Sun system (not to scale) (After Bely 2003)

in the theoretical three-body system, but in the actual solar system, the gravitational pull of the outer planets overwhelms the shallow potential wells at those locations.)

JWST will orbit about L2 using periodic station-keeping maneuvers. It will take the observatory about 3 months to reach its intended position 1.5 million kilometers from Earth. The Sun, Earth, and Moon are always located on the same side of the telescope, and the orbit is chosen so that JWST is never in the shadow of the Earth. The environment is thermally stable and simplifies the design of passive cooling systems. The continuous solar illumination means that the solar panels provide a steady supply of onboard power. The maximum observing efficiency achievable in this orbit is higher than in LEO because there are no periodic occultations of objects by the Earth (Sabelhaus et al. 2005). JWST will also use NASA's Deep Space Network for communications.

4.4 Orbits and Communications for Future Missions

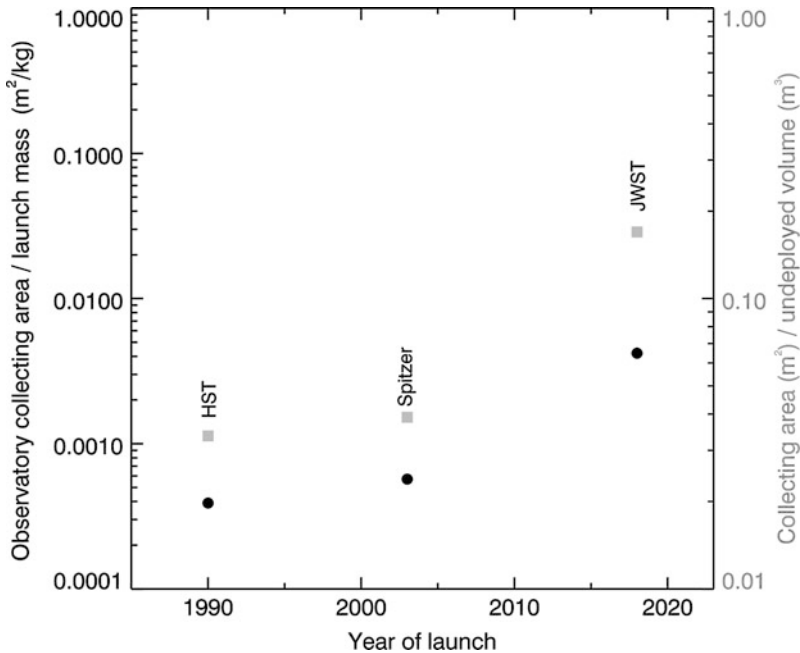
It is likely that future missions will continue to make use of L2 in the Earth-Sun system because of its stable thermal environment and relatively close proximity to Earth. For missions that require lower launch costs or are considering human servicing, a Sun-synchronous low earth orbit might be attractive. In a Sun-synchronous LEO, the satellite orbits about the Earth in the plane perpendicular to the Earth-Sun line, passing over both poles of the Earth. The thermal load from the Sun is constant in such an orbit, leading to better thermal stability than in a low-inclination LEO. The IR load from the Earth varies just as it does for a standard LEO. For additional thermal stability, the telescope pointing can be constrained to face away from the Sun so that it never directly views the IR load from the Earth; a full year is then needed to cover the whole sky.

To enable human servicing, it might be possible to transfer an observatory from an operational orbit far from Earth to a "servicing" orbit that passes close to Earth; the observatory would be returned to its original orbit after the human servicing was complete (Thronson et al. 2005).

JWST will download about 230 Gbits/day of data using a 60 cm diameter high-gain antenna to transmit data to the 34-m Deep Space Network antennas. The proposed ATLAST mission would require a data volume of 200 GBytes/day, about seven times larger than that of JWST, due to its larger detectors. This could be accomplished with a transmission antenna of 120 cm in diameter. (A larger antenna increases the S/N, allowing for the transmission of higher-frequency signals.) However, the antenna pointing requirement would be more stringent since larger antennae have smaller beam sizes. This might require an antenna alignment procedure that would be performed at the start of each downlink pass. Alternate possibilities include developing a higher-power transmitter, longer downlink times to the DSN, the addition of new receive-only antennas to the DSN, using a phased array so that pointing could be adjusted without physical motion of the antenna, or using optical laser communication (Postman 2009).

5 Packaging and Launch Vehicles

Space telescope missions are limited by the volume available in the launch vehicle and by the payload mass that the vehicle can place into the correct orbit. To launch larger primary mirrors, observatories have been forced to become more lightweight and more compact in their launch



■ Fig. 9-23

Over time, increasingly large primary mirrors can be fit into smaller masses and volumes

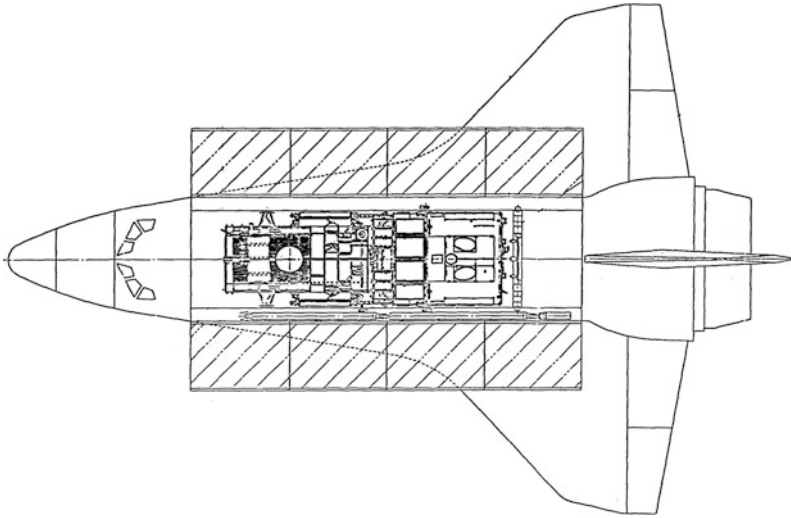
■ Table 9-9

Physical parameters for HST, SST, and JWST. The listed volumes represent an envelope around the maximum dimension of the undeveloped observatory. JWST's collecting area was calculated as the sum of 18 regular hexagonal segments with a flat-to-flat diameter of 1.3 m. JWST's currently planned launch date is 2018

	HST	SST	JWST
Primary mirror diameter (m)	2.4	0.85	6.6
Collecting area (m ²)	4.24	0.49	26.3
Observatory mass (kg)	10,760	860	6,300
Approximate stowed volume of the observatory (m ³)	~13.1 × 4.3 m dia. (190 m ³)	~4 × 2 m dia. (12.6 m ³)	~10.7 × 4.3 m dia. (155 m ³)
Launch Year	1990	2003	2018

configurations. The primary mirror size as a function of mass and volume of the observatory has increased over time, as shown in [Fig. 9-23](#). The values used to calculate the plotted ratios are listed in [Table 9-9](#).

To fit ever-larger collecting areas into the mass and volume provided by the launch vehicles, deployable technologies must be used, or on orbit assembly must be carried out. Most automatic deployments add risk since an unsuccessful deployment could compromise the mission. Everywhere possible, redundant systems are used so that if one deployment method fails, an independent backup method can be used.



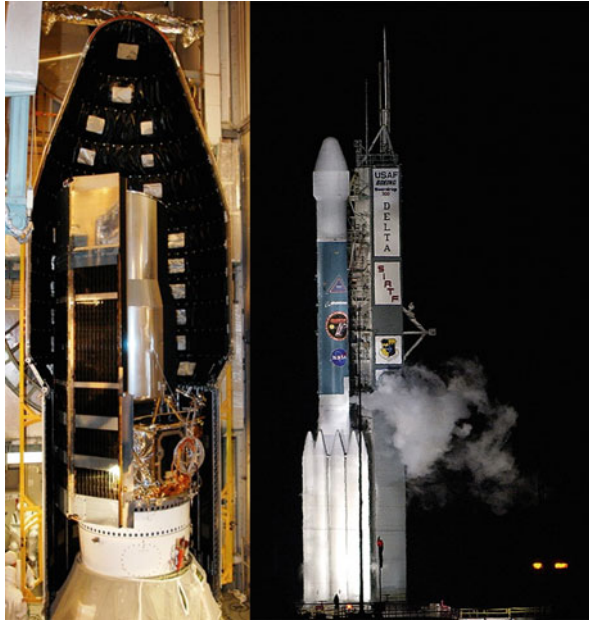
■ Fig. 9-24
Schematic of HST folded into the Shuttle Orbiter for launch (Lallo 2012)

A schematic of the HST in its launch configuration is shown in ● Fig. 9-24. HST's launch mass was approximately 10,760 kg and it was contained inside a volume 13 m long and 4.3 m in diameter. Note that the diameter of the shuttle bay limited the diameter of HST's primary mirror.

HST had three key deployments: the solar arrays, the high-gain antennas, and the aperture door. The solar arrays were rolled into cylinders for launch, and unfurled to their full 2.4×12.2 m size after launch. The two high-gain antennas and their masts were latched to the side of HST during launch and deployed by folding away from the outer shell. The aperture door was closed and latched during launch and opened toward the Sun after launch. The door was designed to open and close multiple times; if the Sun sensors on the spacecraft indicate that the telescope is pointing near the Sun, the door is designed to close automatically. All HST deployments included redundant electronics, bearings, and latches where possible. Because HST is in a low Earth orbit, additional reliability was built into the system by designing all deployable components to allow for manual operation by astronauts.

A photograph of SST in its Delta 7920H fairing is shown in ● Fig. 9-25. The volume occupied is roughly 4 m in length and 2 m in diameter; the observatory's mass is only 861 kg. SST did not use deployable components to reduce its launch volume. Like HST, the diameter of SST's primary mirror is the largest allowed by the chosen launch vehicle. After launch, three critical deployments occurred: a helium vent valve opened, the dust cover at the end of the outer shell was ejected 5 days after launch, and a valve opened the instrument chamber to space (Young 2011, private communication). The limited number of critical deployments results in lower risk of mission failure and a lower overall mission cost, since actuators, controllers, and electronics are kept to a minimum.

JWST launches in an Ariane 5 fairing. The undeployed telescope occupies a volume approximately 10.7 m long and 4.3 m in diameter. After deployment, the observatory's envelope is about $21 \times 12 \times 13$ m. To reach its relatively small deployment volume, multiple parts of JWST deploy after launch.



■ Fig. 9-25

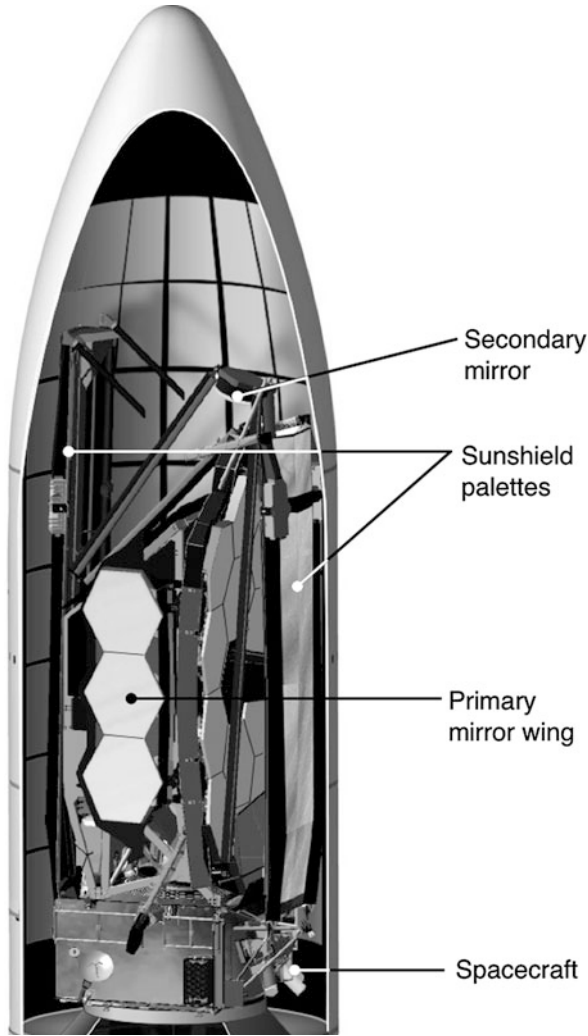
The Spitzer Space Telescope in its Delta 7920H launch shroud and at launch (NASA 2003)

The sunshield is stowed on two vertical palettes for launch, as shown in [Fig. 9-26](#). It deploys to its final area of about 170 m^2 . A “momentum trim flap” deploys from the end of the sunshield. The solar panel (underneath the sunshield) is folded into five segments that unfold into a single plane after launch. A tower that separates the telescope and instruments from the sunshield and spacecraft also deploys.

The telescope portion of the observatory deploys in three steps: the secondary mirror’s support structure unfolds from its vertical position. Then, the left and right primary mirror “wings” fold forward to create the 18-segment primary mirror. Finally, the primary mirror segments and secondary mirror are actuated out of their stowed launch position to a point where the optical alignment process can begin (Lightsey et al. 2012). (Alignment of the optics and primary mirror segments are discussed in [Sect. 6.4](#)).

In the near term, missions will continue to be limited by the mass and volume constraints of the available launch vehicles, and missions that require aperture diameters larger than the available fairings will continue to rely on deployable primary mirrors and secondary mirror supports. Currently, NASA is developing a heavy-lift launch vehicle known as the Space Launch System (SLS), which will initially be capable of putting 70 metric tons into LEO, and will ultimately reach 130 metric tons. The fairing diameter for cargo could be as large as 8 m in diameter, which would allow larger observatories to be carried to L2.

In the long term, it may be possible to bypass the constraints imposed by launch vehicles by assembling observatories in space. The observatory could be divided into segments that could launch separately and would autonomously assemble in orbit. Observatories could also potentially be assembled by astronauts in LEO or in depots located further away from the Earth (Goddard Space Flight Center 2010).



■ Fig. 9-26

A graphic of JWST stowed for launch in the Ariane 5 fairing. The secondary folds upward, the primary mirror “wings” fold back around the instrument radiators and compartment, and the sunshield folds into two palettes at the front and back of the assembly (NASA 2003, labels added)

6 Optical Considerations

Multiple requirements affect the optical design and fabrication of a space telescope. The optical system must achieve diffraction-limited imaging below a specified wavelength and provide a field of view large enough to accommodate the onboard instruments. Pupils and image conjugates might be required at certain locations so that they can be used for stray light control and for pointing adjustments. The technology used to fabricate the mirrors must be lightweight

so that the telescope stays within its mass budget. The alignment of the optical system must be maintained, either through structural stability or through active control of the mirrors once on orbit.

6.1 Optical Designs and FOV Allocations

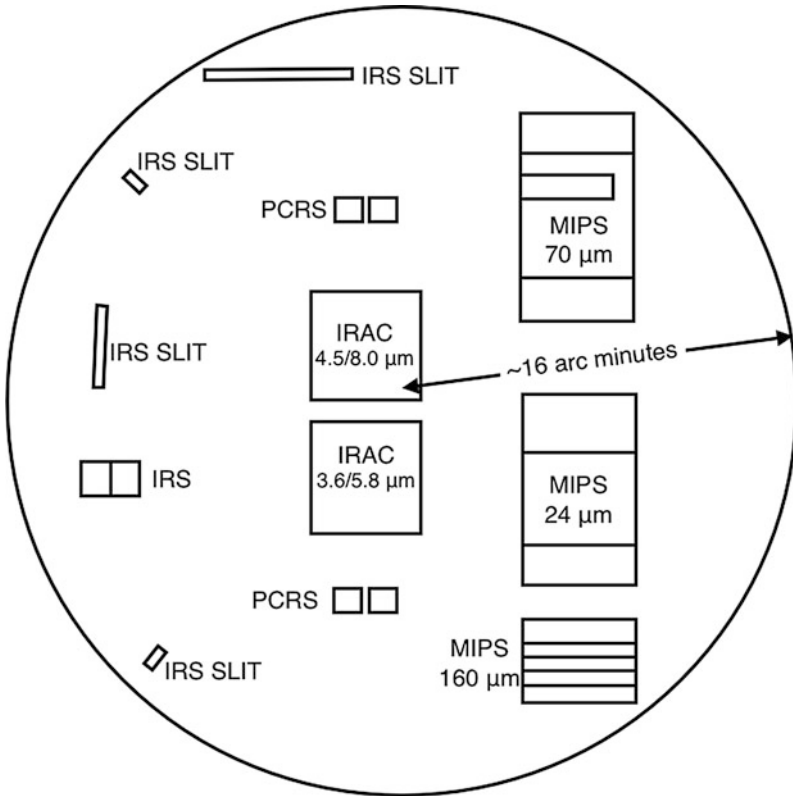
HST and SST are both variations on standard Cassegrain telescopes, which are compact two-mirror telescopes with convex secondary mirrors. Cassegrains, with convex secondaries, are more compact than systems with concave secondaries which helps to minimize a system's launch volume. (The convex secondary is more difficult to fabricate and test, however, since test beams diverge from the surface and require very large relay or collection optics.) Two-mirror systems also minimize the number of reflections within the OTA, which is especially important for HST since it is difficult to achieve high reflectivities in ultraviolet wavelengths. Each reflection off of a typical UV-coated mirror results in a 10% loss of intensity relative to the incident beam. Also, current detector quantum efficiency (QE) is low (typically 10–20%), so a 10% loss from an extra mirror reflection represents a significant degradation of science capability.

Specifically, HST and SST both use Ritchey-Chretien designs in which spherical aberration and coma are both corrected, and the aberration that limits the field of view in these systems is astigmatism (Schroeder 1987). Both HST and SST have large fields of view in which instruments can be placed: 32 arcmin in the case of SST and 28 arcmin in the case of HST (Lee et al. 1998). HST achieves diffraction-limited imaging for wavelengths above about $0.6\ \mu\text{m}$, while SST is diffraction-limited beyond $5.5\ \mu\text{m}$.

The instrument locations within the available fields of view for HST and SST are shown in [Fig. 9-27](#) and [9-28](#). Since the instruments are separated in field space, the observatory is repointed to select a particular instrument. Observations with more than one instrument (parallel observations) can be made if the telescope is undertaking large areas surveys, or simply by exploiting the “serendipity” of observing coincident areas of the sky next to targets of interest. Within instruments, light may be split into different wavelength regimes so that objects can be observed simultaneously in multiple wavelengths.

JWST is a three-mirror anastigmat (TMA), as sketched below in [Fig. 9-29](#). The aberrations that limit a telescope's field of view scale linearly with aperture diameter. JWST, with its 6.6-m primary mirror, uses three conic mirrors to achieve its 18×9 arcmin field of view. The approximate locations of the instruments within that FOV are shown in [Fig. 9-30](#). The gold-coated mirrors have high reflectivities in the IR and the QE of the IR detectors is high, so the throughput of the system remains high despite the additional reflections within the OTA. JWST is required to achieve diffraction-limited imaging for wavelengths above $2\ \mu\text{m}$, but its performance may be better than required.

An advantage of TMA systems is that they provide a real pupil within the optical train. The pupil is an image of the primary mirror, and that image does not change as a function of the object's position in the sky or as a function of wavelength. A planar mirror located at a pupil plane can be tilted to steer all objects in the field of view by the same amount. The mirror can refine the pointing of the observatory and may also be used to compensate for jitter and other disturbances in the structure of the observatory. As sketched in [Fig. 9-29](#), JWST's FSM mirror occurs at a real pupil and is used to make fine adjustments to the observatory's pointing.



■ Fig. 9-27

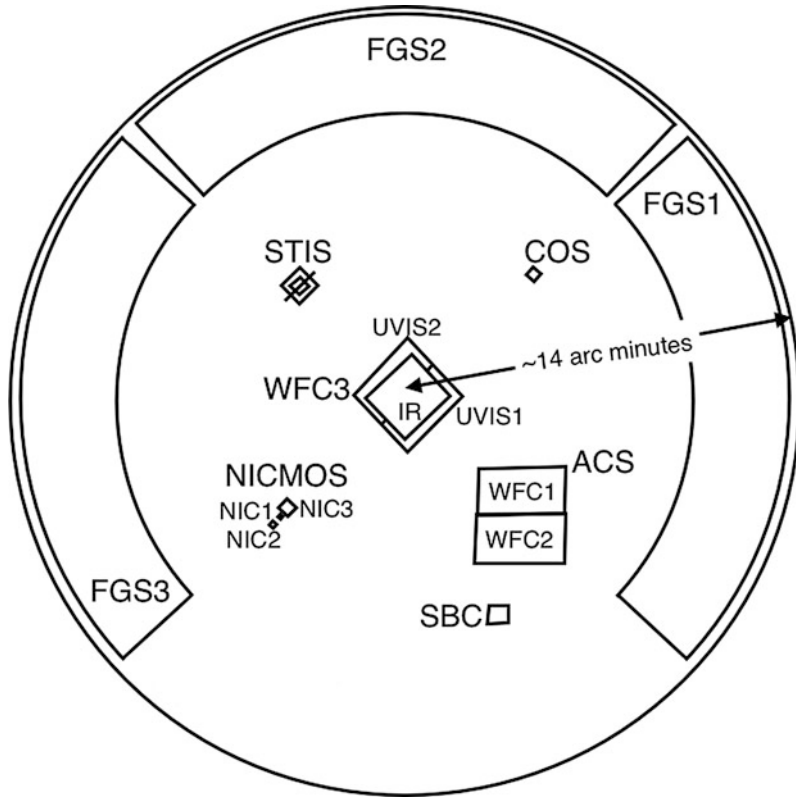
Field allocations for the instruments aboard SST. The field allocations for the MIPS detectors include the range accessible with the instrument's scanning mirror (After Spitzer Science Center 2011)

6.2 Stray Light Control

Stray light is defined as any photon that reaches an instrument's detector that did not come from within the science field of view or that arrived at the detectors via an unplanned path. Sources of stray light for orbiting observatories include the Sun, Earth, Moon, other celestial bodies, galactic and extragalactic sources, zodiacal light, and in infrared wavelengths, self-emission from the observatory itself.

OTA enclosures are a primary defense against stray light from a broad range of sources. Field stops, placed in planes optically conjugate to the detectors, block stray light from celestial objects outside of the detector's field of view. Other baffles may be used at key locations inside the instruments. Instruments and instrument compartments are also enclosed.

The telescope sections of both SST and HST are enclosed in outer shrouds. In [Fig. 9-39](#), some of the elements used for stray light control in HST can be seen: the open aperture door on the sunward side of the observatory prevents sunlight from entering the shroud directly. The inner surface of the door is covered with black paint to reduce its reflectivity in the visible. Inside the shroud, opposite the aperture door, a series of black baffles further prevents stray

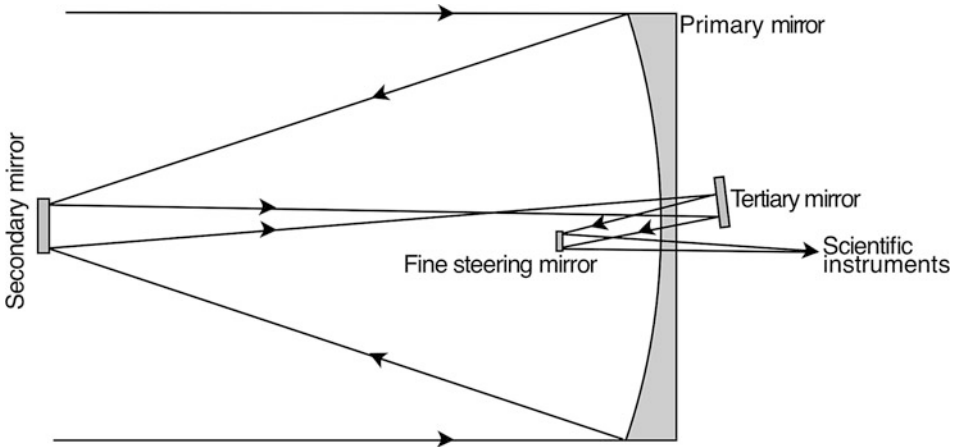


■ Fig. 9-28

Field allocations for the instruments aboard the Hubble Space Telescope after Servicing Mission 4 (After Space Telescope Science Institute 2010)

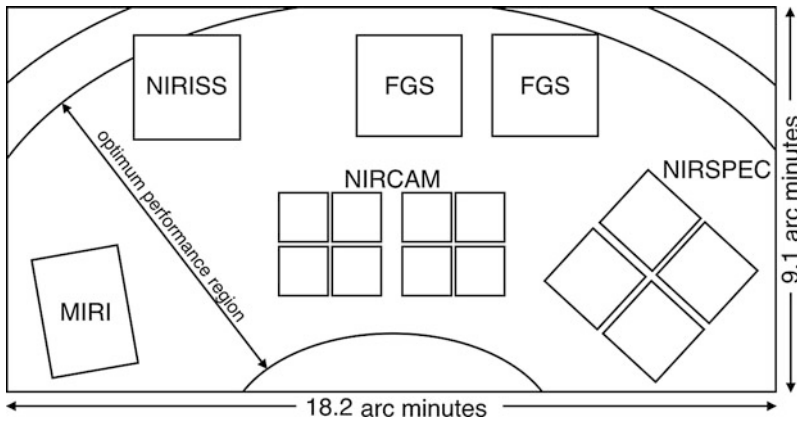
light from glinting off the inside of the shroud. All materials that cover the shroud have inner surfaces that are black and absorb visible wavelengths.

For JWST, an OTA enclosure was omitted because its large size would have added substantial mass and would have been difficult to deploy and cool. Instead, a sunshield is used to protect JWST's optics from stray light and thermal effects from the Sun, Earth, and Moon, which are always located on the same side of the observatory (see ● Fig. 9-22). The accessible pupil within JWST's TMA design is also important for stray light control, and JWST's FSM carries a pupil mask. Since the image of the primary mirror at a pupil is constant as a function of wavelength and field angle, a mask that lies just outside the edges of the primary image will block any photons coming from objects just outside the primary mirror. (Such photons would miss the primary mirror, but fall directly onto the secondary mirror and make their way to the instruments.) The mask also blocks the central obscuration of the telescope so that photons cannot pass the edges of the secondary mirror and fall directly onto the instrument pickoff mirrors (Lightsey et al. 2012). Another layer of stray light control for JWST is a field stop placed near the Cassegrain focus of the system, at the end of the AOS.



■ Fig. 9-29

A schematic of the JWST optical design, which is a three-mirror anastigmat. The fine steering mirror is located at a real pupil and can be steered to refine the observatory's pointing

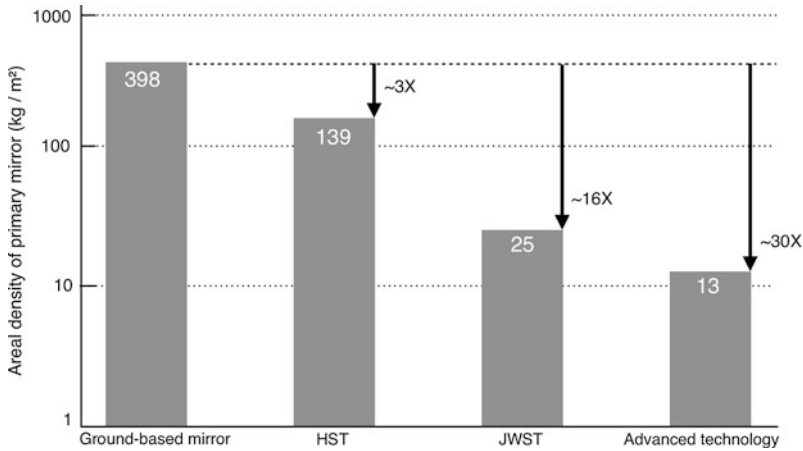


■ Fig. 9-30

A sketch of the JWST field of view, as projected onto the sky. The allocations for each instrument's field of view are also shown. (Not shown: long-wave NIRCam, NIRCam coronagraph, MIRI coronagraph, NIRSpec fixed slits, imaging window or IFU field points, and MIRI mid-resolution spectrograph FOV)

6.3 Lightweight Mirror Technologies

The total mass of a space-based observatory is strictly limited by the launch vehicle, and virtually every component is lightweighted compared to components used in ground-based systems. These lightweight structures must survive the rigors of launch and operate in the zero-gravity on orbit environment, despite that fact that they are fabricated, aligned, and tested in a 1-g environment. Primary mirrors are obvious targets for lightweighting since they are often the



■ Fig. 9-31

A comparison of the areal densities of various primary mirrors shows a significant decrease over time. HST used a lightweighted glass mirror, JWST and SST used a beryllium technology, and technologies for future missions are achieving even smaller areal densities (Postman 2011, private communication)

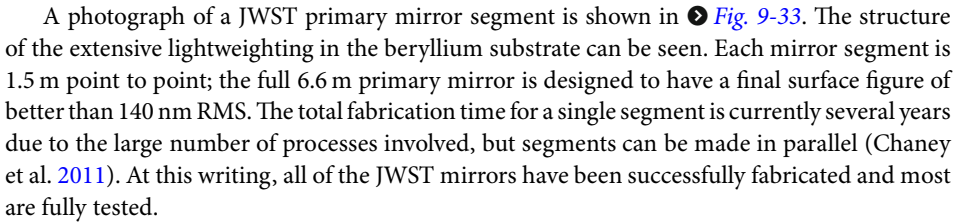
largest single element in an observatory. HST has a lightweighted glass primary that comprises about 6% of HST's total mass and has an areal density of 140 kg/m².

JWST and SST, more tolerant of wavefront error because of their longer wavelength ranges, took advantage of an actuated beryllium mirror technology to reach areal densities of approximately 25 kg/m² (including the actuators and support structures). Future missions, with ever-larger primary mirrors, will require even lower areal densities. A comparison of the areal densities of primary mirrors is shown in [Fig. 9-31](#).

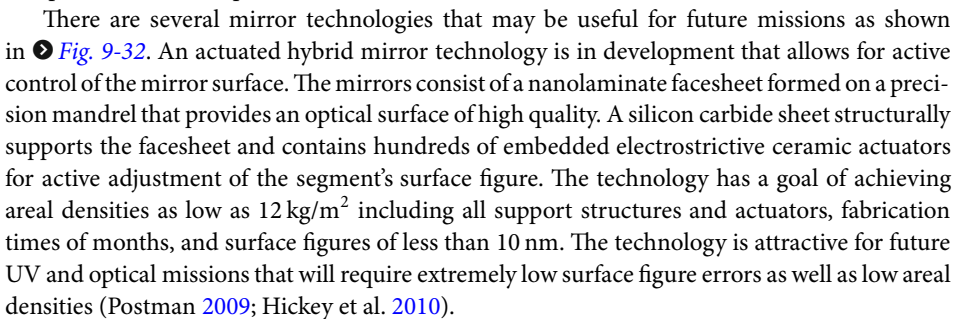
HST's primary mirror is fairly rigid and consists of a lightweighted core with front, back, and edge plates fused to it. The OTA is estimated to contribute about 18 nm of surface error (mid- and high-spatial-frequency polishing errors) to the system, and the primary mirror was polished to about 10 nm RMS. (This estimate excludes the low-spatial-frequency spherical aberration that was found after launch, and that is now corrected within each Hubble instrument (Krist and Burrows 1995)).

Both SST and JWST use lightweighted beryllium optics. The fabrication process begins with the machining and etching of a beryllium blank to produce an extremely lightweight substrate. The mirrors are initially polished to have a good surface figure at an ambient temperature. The mirrors are then taken to cryogenic temperatures (5.5 K for SST, 40 K for JWST) and the surface figures measured. The negative of the cryogenic surface errors are then polished into the mirror so that when the cryogenic surface figure test is repeated, the mirrors are nearly perfect (Lee et al. 1998).

JWST's 18 segments are semirigid; they can be adjusted in six rigid-body degrees of freedom and in radius of curvature so that they can be aligned to perform as a single primary mirror. The quality of the surface figure depends largely on the quality of the final polish and the rigidity present in the beryllium structure. Individual segments are required to reach surface figure errors of better than 30 nm RMS.

A photograph of a JWST primary mirror segment is shown in  Fig. 9-33. The structure of the extensive lightweighting in the beryllium substrate can be seen. Each mirror segment is 1.5 m point to point; the full 6.6 m primary mirror is designed to have a final surface figure of better than 140 nm RMS. The total fabrication time for a single segment is currently several years due to the large number of processes involved, but segments can be made in parallel (Chaney et al. 2011). At this writing, all of the JWST mirrors have been successfully fabricated and most are fully tested.

For future missions with larger primary mirrors, lighter mirror segments will be required. Such segments are likely to have less rigidity and require active control of the surface figure of each segment to achieve low figure errors. This is especially true for future optical and UV missions; the ATLAST mission is specified to be diffraction-limited at a wavelength of 500 nm, which would require mirror surfaces of high quality (~ 10 nm RMS surface errors to achieve a total WFE budget of ~ 36 nm). While the required surface errors are comparable with those achieved on the HST primary, the ATLAST mirrors need to achieve those errors on mirrors with areal densities ten times less than HST's. Because of the demanding surface error requirements, one option proposed for ATLAST is an 8-m monolithic primary that would fit into a future heavy-lift launch vehicle. Segmented mirrors are also being considered and may be more compatible with the capabilities of near-future launch vehicles.

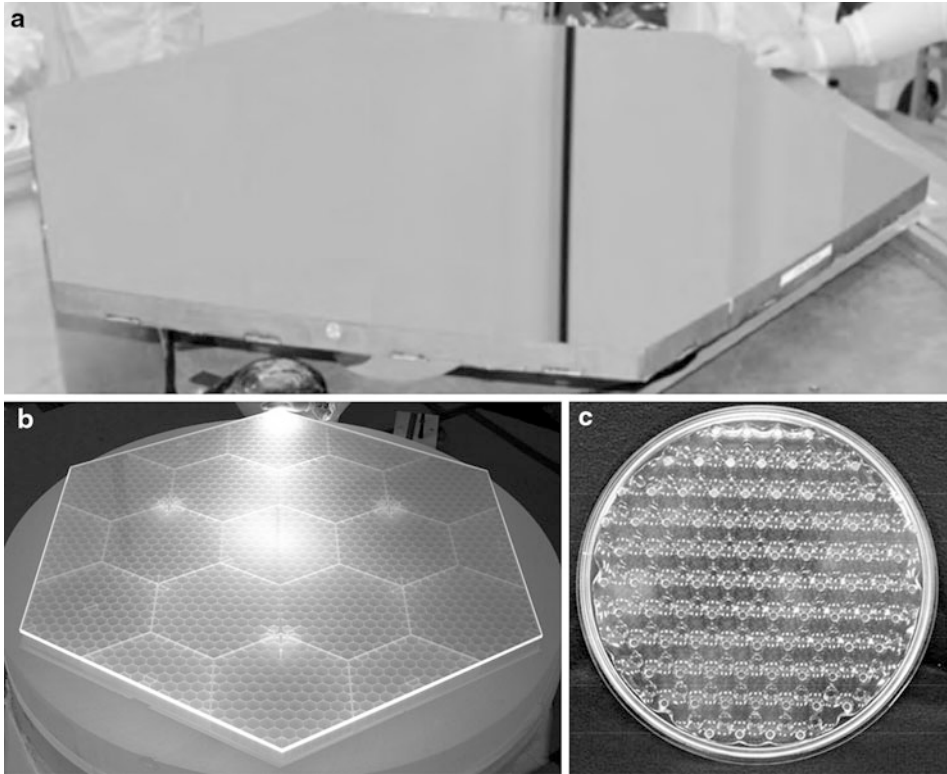
There are several mirror technologies that may be useful for future missions as shown in  Fig. 9-32. An actuated hybrid mirror technology is in development that allows for active control of the mirror surface. The mirrors consist of a nanolaminate facesheet formed on a precision mandrel that provides an optical surface of high quality. A silicon carbide sheet structurally supports the facesheet and contains hundreds of embedded electrostrictive ceramic actuators for active adjustment of the segment's surface figure. The technology has a goal of achieving areal densities as low as 12 kg/m^2 including all support structures and actuators, fabrication times of months, and surface figures of less than 10 nm. The technology is attractive for future UV and optical missions that will require extremely low surface figure errors as well as low areal densities (Postman 2009; Hickey et al. 2010).

An ultralow-expansion (ULE) glass semirigid mirror technology was initially developed with a goal of an areal density of 15 kg/m^2 . Faceplates over a honeycomb structure are fabricated in a planar configuration, then “slumped” over a precision mandrel to achieve off-axis and aspheric shapes. Standard polishing techniques are used to finish the mirrors. The mirrors alone achieve an areal density of 8 kg/m^2 (Matthews et al. 2003).

A “replicated corrugated” borosilicate mirror technology achieves areal densities of 11 kg/m^2 (not including mounting hardware or any actuation). The mirrors are fabricated from flat glass sheets and consist of five layers: a front facesheet, a “microcore,” an inner sheet, a “macrocore,” and a back facesheet. The mirrors are replicated on precision surfaces so that minimal polishing is required to finish the mirrors. The technology is especially attractive for systems requiring a large number of individual mirror segments because of its low cost and short fabrication times (Egerman et al. 2010; Strafford et al. 2006).

6.4 Optical Alignment

HST was aligned on the ground and largely relied on structural stability to keep the optical aberrations low through launch and into orbit. To compensate for small structural changes, though, the secondary mirror is actuated so that the telescope focus can be controlled. Focus



■ Fig. 9-32

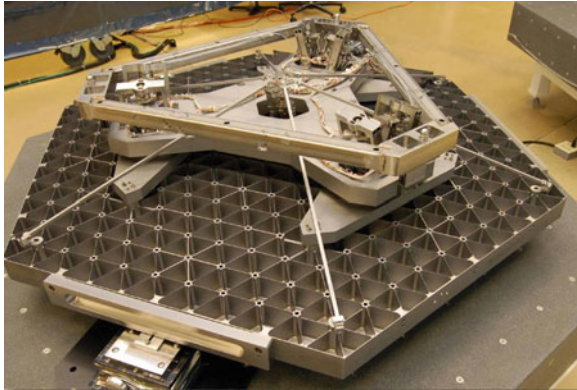
Lightweight mirror technologies under development include: (a) actuated hybrid mirrors (Hickey et al. 2010), (b) fused ULE mirrors (Matthews et al. 2003), and (c) replicated corrugated borosilicate mirrors (Egerman et al. 2010)

adjustments have been necessary once or twice a year, at most.⁵ In addition, most of HST's instruments have focus adjustment mechanisms to compensate for small amounts of focus error. HST's primary mirror does have 24 actuators that allow for limited correction of some low-order aberrations in the primary mirror, but they have not been used since commissioning. (Their range is too small to correct the spherical aberration that was detected in the primary mirror early in the mission. Corrective optics that compensate for the aberration have been incorporated in the later generations of science instruments (Lallo 2012)).

Similarly, SST has an adjustable secondary to compensate for small structural changes on orbit. Its instruments do not include focus adjustment mechanisms, though, and the primary is not actuated. Without focus adjustments, the instruments must be designed to tolerate a larger amount of defocus, or the structural changes in the system must be held to low levels. The advantage of the approach is that the number of actuators in the system is minimized, reducing risk, complexity, and power consumption.

A more complex alignment approach was required for JWST due to its segmented primary mirror and the addition of tertiary and FSM mirrors. Optical alignment has to occur on

⁵<http://www.stsci.edu/hst/observatory/focus/mirrormoves.html>



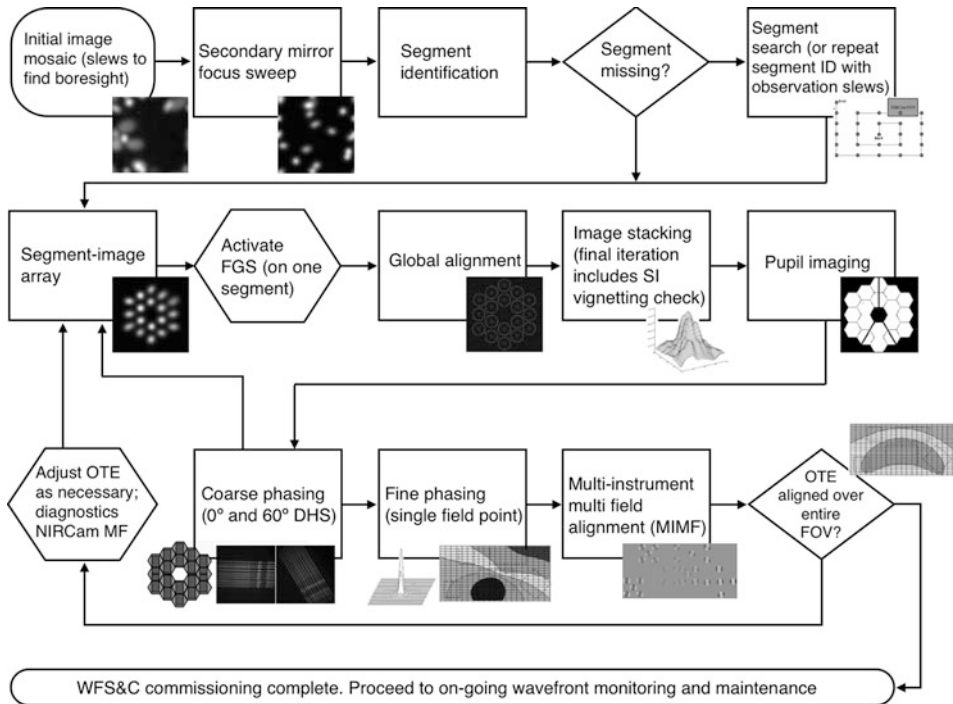
■ Fig. 9-33

A photo of the back of a JWST primary mirror segment. Extensive lightweighting of the beryllium mirror can be seen. Seven actuators provide rigid-body motions and a radius of curvature adjustment (NASA 2010)

orbit after the deployment of the segments and the secondary mirror. The alignment method is image-based, and relies primarily on images from the NIRC*am* instrument. Initial alignment of the telescope is expected to take several weeks. For maintenance of the optical alignment, the telescope could be adjusted as frequently as every 2 weeks.

To accomplish the on orbit alignment, most of JWST's mirrors are active. For each of the 18 primary mirror segments, six actuators in a hexapod configuration provide 6 degrees of freedom (DOF) of rigid-body motion as shown in ● Fig. 9-33. A seventh actuator and a set of struts provide a radius of curvature adjustment. The secondary mirror also includes a hexapod for 6 DOF rigid-body motions. The FSM, which is a planar mirror, can be moved in tip and tilt. The tertiary mirror position is fixed. In addition to the active mirrors in the telescope, all of JWST's NIR instruments contain focus adjustment mechanisms. Each of the two redundant NIRC*am* modules can also align their internal apertures to the telescope exit pupil using tip/tilt motions of their focus adjust mechanisms.

A wavefront-sensing and control (WFSC) process has been developed to control on orbit alignment of JWST's mirrors. The process is fairly complex and requires human oversight; a schematic of the full process is shown in ● Fig. 9-34. The telescope is pointed toward a bright, isolated star. When the primary mirror segments are initially deployed, they have large tilt, piston, and other errors with respect to one another. Each segment produces a separate image of the star on the image plane. The first few steps of the alignment process put the secondary in an average focus position, identify which image belongs to which segment, and locate any segment images that are not found in the NIRC*am* image plane. The segment images are then driven into an array on the image plane, and “global alignment” optimizes the position of each segment to produce the best image possible. The images are then stacked at the center of the image plane, and the pistons are corrected using the coarse phasing process. The telescope is now sufficiently aligned to carry out fine phasing, which is a phase retrieval process (Gerchberg and Saxton 1972) that completes the alignment process. Finally, the “multi-instrument multifield” (MIMF) step checks the performance of the telescope at the other instruments and the process is repeated if necessary. After this initial alignment process, the fine-phasing step will be repeated every 2 weeks as necessary to maintain the alignment of the mirrors (Lightsey et al. 2012).



■ Fig. 9-34

A schematic of the JWST commissioning process that is used to align the telescope (Lightsey et al. 2012)

6.5 Optical Alignment for Future Missions

HST and SST relied on structural stability and infrequent focus adjustments of the secondary mirrors to maintain optical alignment while on orbit. JWST, with lower areal density and less structural stability, will rely on its active primary mirrors segments, secondary mirror, and FSM to periodically correct its alignment. As primary mirrors in future space telescopes get larger and the areal densities of the observatories fall, the systems are likely to have still less intrinsic structural stability. To maintain the alignment and optical surface figures of these systems, the control philosophy used will have to more closely resemble that of fully active and adaptive ground-based telescopes.

In fact, the evolution of space-based systems is beginning to parallel that of ground-based telescopes, which have gone from passive optical alignment (using the stability of the telescope structure to maintain alignment), to actively controlled telescopes like the segmented 10 m Keck Telescopes, and finally to completely active and adaptively controlled telescopes such as the European Very Large Telescopes, the two 8-m Gemini Telescopes, and all the future 20–40-m ground-based telescopes now in the planning stages. 📌 Table 9-10 illustrates this evolution in optical control philosophy.

JWST must steer away to a bright isolated star each time the alignment is adjusted, and a human operator must oversee the alignment process. Future space observatories will have to carry out alignment more frequently and might require fully active adjustment of mirror surface figures. The WFS&C process for these systems will have to be fully automated, and the

■ Table 9-10

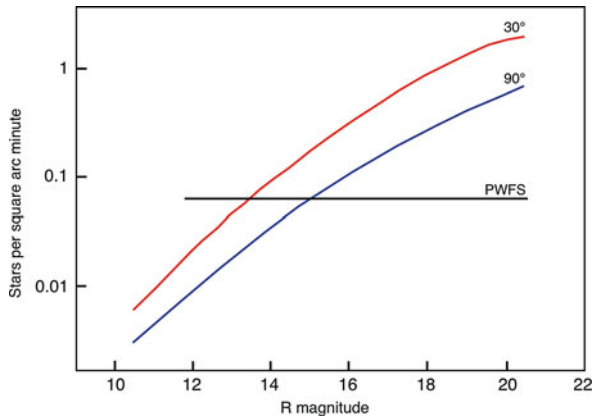
The evolution of optical and structural control philosophies for increasingly larger telescopes, both on the ground and in space

	Passive	Active	Active and adaptive	Fully adaptive
Optical alignment and optical figure control approach	Relies on intrinsic structural stability and occasional focus adjustments with secondary using a reference star	Moderate bandwidth (>1 Hz) internally referenced alignment with infrequent (>week) calibration of alignment and phasing of primary mirror using a reference star	Low bandwidth (<0.01 Hz) active correction of alignment and optical figure using a reference star, with high bandwidth (>100 Hz) adaptive tip/tilt and focus correction using a guide star	Moderate to high bandwidth (>10 Hz) continuous correction of alignment, optical figure and phasing using a guide star
Ground-based telescopes	<ul style="list-style-type: none"> • Palomar (5 m) • KPNO (4 m) 	<ul style="list-style-type: none"> • UKIRT (3.8 m) • UofA MMT • Keck (10 m) 	<ul style="list-style-type: none"> • VLT (8 m, 4 ea.) • Gemini (8 m, 2 ea.) • Magellan (6 m, 2 ea.) 	<ul style="list-style-type: none"> • GMT (20 m) • E-ELT (40 m)
Space-based telescopes	<ul style="list-style-type: none"> • HST (2.4 m) • Spitzer (0.85 m) 	<ul style="list-style-type: none"> • JWST (6.6 m maximum diameter) 		<ul style="list-style-type: none"> • ATLAST concept (10 m)

adjustments will have to be made without steering away from the science target, perhaps by using the images of the guide stars for WFSC operations. From space, telescopes with diameters of 8 m or greater should always be able to find sufficiently bright stars in the periphery of the science field to allow continuous alignment and wavefront control. ▶ [Figure 9-35](#) shows the availability of wavefront-sensing stars for the ground-based 8-m Gemini telescope. Gemini uses wavefront information from a star selected in the periphery of the science fields to update the telescope alignment and the primary and secondary mirror surface figures at a rate of 0.03 Hz. For a space telescope, the number of available stars will be higher due to lower background levels and improved image quality in the absence of atmospheric turbulence.

7 Pointing and Control Systems

Pointing and control systems are needed in order to position the science target at the appropriate location in the observatory's field of view, and to maintain that pointing despite disturbances to the observatory. In general, the rule of thumb used by telescope designers is that the pointing stability should be better than 1/8 the size of the PSF FWHM. External pointing disturbances include solar pressure, and for telescopes in LEO, gravity gradients, and aerodynamic torque. Such disturbances are usually low in frequency. Internal, high frequency disturbances are known as "jitter," and are usually too high-frequency to be controlled using the pointing system. Sources of jitter include any moving or vibrating part in the observatory: reaction wheels, fuel slosh, solar arrays and sunshields, thrusters, gyroscopes, filter wheels, actuators, cryocoolers, and releases of structural strain.



■ Fig. 9-35

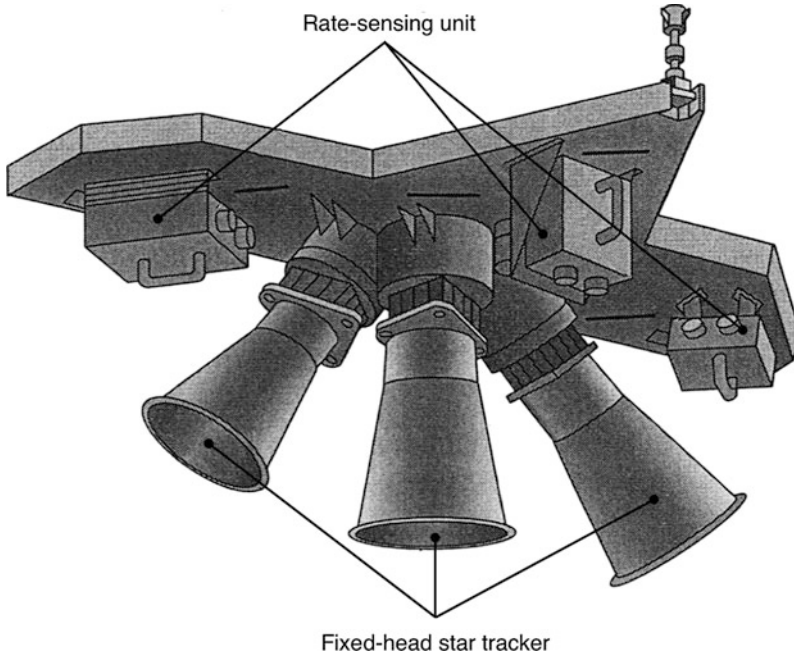
The surface density of stars as function of optical R band magnitude for the galactic latitudes of 30° and 90° from the Galactic plane derived from the Bahcall–Soneira model (1984). The *horizontal line* represents the patrol area of the Gemini Peripheral Wavefront Sensor (PWFS). Where this *line crosses the stellar curve* indicates the magnitude where, statistically, one star should lie within the PWFS patrol area (approximately 14 arcmin in diameter). For example, statistically, one star of magnitude $R = 13.5\text{--}15$ will be available for wavefront sensing for any science field between galactic latitudes of 30° and 90° (<http://www.gemini.edu/sciops>)

Most spacecraft have multiple layers of pointing control. At the first level, a coarse-pointing system uses star trackers and gyroscopes to measure the spacecraft's orientation in space, or attitude.

Star trackers are self-contained cameras that are affixed to the outside of the spacecraft. They have very wide fields of view (about $8 \times 8^\circ$) so that they can compare multiple bright star positions to a star catalog in order to determine the spacecraft's orientation in space. Most missions include three star trackers. The best star trackers have about 10 arcsec RMS absolute attitude accuracy. The location of HST's star trackers are shown in ► Fig. 9-12, and a more detailed sketch of the system is shown in ◉ Fig. 9-36. (The most visible part of the star trackers are the relatively large baffles used to prevent stray light from reaching the tracker optics and detectors.) The star trackers and gyroscopes are mounted in the aft shroud on a dimensionally stable composite bench so that the locations and orientations remain constant. The actual star trackers can be seen in the photograph in ► Fig. 9-37.

A gyroscope is a flywheel that spins at a high rate and senses and measures changes in the orientation of the spin axis (Den Hartog 1961). The flywheel is on a single gimbal and only the corresponding axis is sensed, so three gyroscopes are needed to sense all three angles of a spacecraft's orientation. The rotation of the output axis is proportional to the spacecraft's angular rate about the input axis. That rotation is balanced by electromagnetic torque, and the torque required provides a measure of the spacecraft's angular rate. Most spacecraft use “rate gyros” (Thomson 1986).

Onboard pointing control systems feed this information from the gyroscopes and star trackers into the pointing control law, in order to accurately slew the observatory using the



■ Fig. 9-36

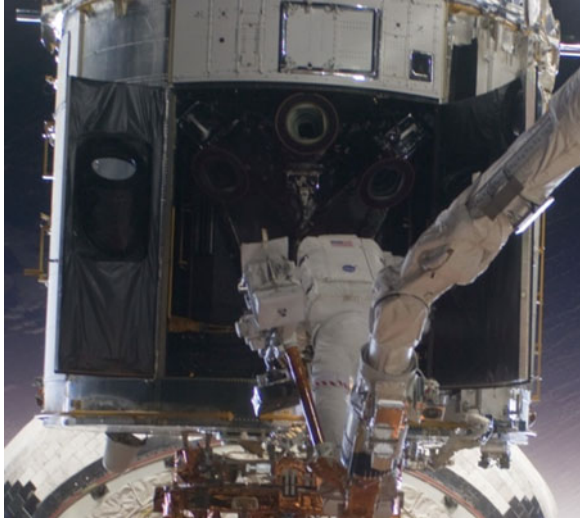
HST carries three star trackers so that the spacecraft's attitude can be determined. Large baffles protect the trackers from stray light from the Sun. Gyroscope units share the support structure with the trackers (After Gilmore 2002)

spacecraft's reaction wheels. Reaction wheels are flywheels that can spin in either direction about their axes up to some maximum rate of several thousand r.p.m. When accelerated or decelerated, the wheels transfer momentum to the observatory since the total momentum of the observatory and reaction wheels must remain constant. Most missions carry four reaction wheels; one for each axis and one spare.

The observatory is not completely isolated from external momentum sources, though. Over time, momentum is added to the observatory system due to radiation pressure and, in LEOs, to gravity gradients. The added momentum can cause the reaction wheels to spin faster until they reach their maximum speeds; control of the observatory would then be lost. The additional momentum must be "dumped" every few hours. HST orbits within the Earth's magnetic field, so momentum desaturation is done continuously using magnetic torquers (Lockheed Missiles and Space Company 1985).

For JWST and SST, momentum desaturation is done periodically using thrusters. SST requires momentum dumping fairly infrequently – about once a week.

Coarse-pointing systems are generally unsuitable for scientific observations since both image stability and absolute pointing accuracy is inadequate due to physical changes in the gyroscopes and to sensing noise in the star trackers. For example, HST's gyros achieve 0.25 mas resolution but drift at a rate of 1 mas/s, while the star tracker performance is rarely better than 10 arcsec. Coarse-pointing accuracy is also limited by the fact that the star trackers and gyroscopes are mounted on the spacecraft bus, and they must be calibrated to match the telescope's



■ Fig. 9-37

Astronaut Michael Good, STS-125 mission specialist, in a session of extra-vehicular activity. Access panels in the aft shroud are open, and the star tracker assemblies are visible (NASA 2009)

boresight. Any changes in the alignment between the spacecraft and telescope (usually due to thermal changes in the observatory structure) cause an error in the coarse-pointing.

Fine-pointing supplements star tracker and gyroscope measurements with data from fine guidance sensors located in the focal plane of the telescope. (The field locations of HST and JWST's FGSs are shown in ► *Figs. 9-28* and ► *9-30*.) After the observatory is pointed toward the science target, the FGS searches for and acquires a particular guide star that has been preselected by the ground system. The FGS aperture size and instrument sensitivity are designed to provide access to a suitable guide star for at least 95% of all telescope pointings at a given roll about the boresight. Once the guide star image is located by the FGS, its position is used by the Attitude Control System (ACS) to refine the pointing and stabilize the observatory to place the science target image at the correct location in a science aperture.

For HST, three FGSs encircle the outer perimeter of the field of view as shown in ► *Fig. 9-28*. Each guider consists of two sets of scanning mirrors (star selector A, star selector B), a polarizing beam splitter (creating two beams with orthogonal polarizations), an X-axis and Y-axis shearing interferometer, and two photomultiplier tubes (PMTs) on each axis. The system has an instantaneous field of view (IFOV) of 5×5 arcsec, which can be placed anywhere in the full 69 arcmin^2 FOV using the star selectors. At the end of the spacecraft slew to the science target, the FGS IFOV is commanded to the expected location of the guide star. The FGS flight software commands the instrument to execute a spiral search for the guide star, which is detected using the PMTs. Once located, the FGS IFOV nutates about the location of the guide star's image to precisely determine the image's photo-center. Finally, the FGS interferometrically acquires the guide star in "fine lock" by using the PMT data to locate and lock on to (track) the null point of the sheared interference fringe. The location of the guide star in the FGS FOV is updated and reported every 25 ms.

On HST, two FGSs are used simultaneously. One provides data that is used for pitch/yaw control, while the other provides guide star centroids used for roll control. Guide star magnitudes can range from $9 < V < 14$. Fine guidance on HST achieves better than 7 mas image stability. Absolute pointing accuracy is limited primarily by the accuracy of the guide star catalog position on the celestial sphere (the error is expected to be ~ 0.3 arcsec (Space Telescope Science Institute 2010)).

The JWST FGS is a cryogenic instrument with two independent $2k \times 2k$ NIR detectors covering 2.3×2.3 arcmin on the sky (see [Fig. 9-30](#)). During a science observation, only one channel is used and the other is idle. At the end of the spacecraft slew to the science target, the FGS is commanded to locate the guide star in a particular channel. It does so by obtaining a “full frame” image of the channel’s entire FOV. It then compares the observed scene to the predicted scene provided by the ground system, and applies a pattern match to locate the guide star image.

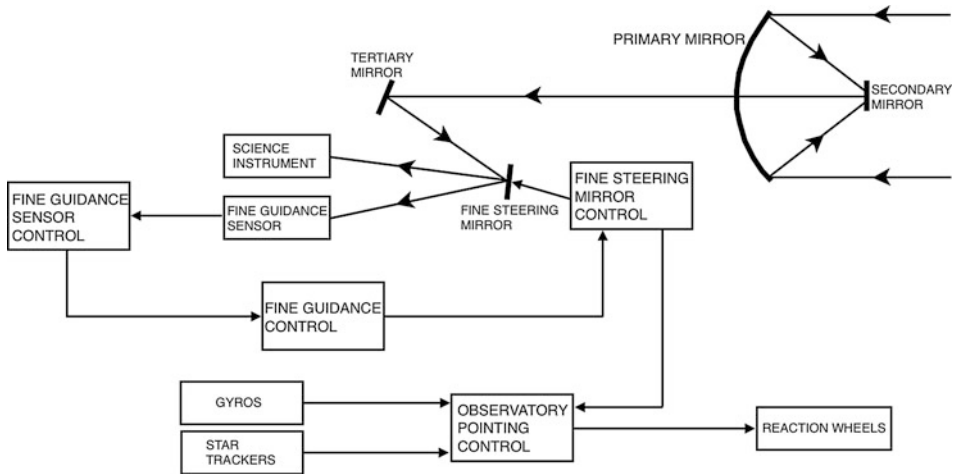
After identifying the guide star, the FGS places a subarray on the guide star, computes the centroids, and reports them to the ACS to improve pointing accuracy. The process is repeated with smaller subarrays to provide increasing centroiding accuracy. Once the guide star is within 3 arcsec of its desired location, the FGS continuously “tracks” the guide star using a smaller subarray that can be frequently repositioned to follow the guide star. At this point, the ACS enters its Fine Guide Control, using the FGS data to drive the FSM in a closed loop, as sketched in [Fig. 9-38](#). The FSM has a finite range, so adjustments to the observatory pointing via the reaction wheels are also made to keep the FSM near the center of its range. When the star is within 35 mas of its desired location, the FGS images the guide star with an 8×8 pixel subarray at a rapid rate.

Note that the FGS data are used only for pitch/yaw control. JWST roll is controlled by the star trackers. JWST will use guide stars as faint as $J < 18.5$. Using the FGS, star trackers, and gyros, JWST is expected to achieve 0.007 arcsec pointing stability for long-duration science exposures. As with HST, the absolute pointing accuracy is limited primarily by the accuracy of the guide star’s cataloged position on the celestial sphere (the error is expected to be ~ 0.3 arcsec).

SST does not use a fine guider that provides real-time pointing measurements. Rather, it uses two small sensors instruments (labeled “PCRS” in [Fig. 9-27](#)) that can be used to check the relative alignment between the (cold) telescope boresight and the (warm) spacecraft’s coarse-pointing system a few times per day. A calibration is done by placing a catalog star on the PCRS and comparing its position to the positions of other catalog stars in the star trackers. Roll can also be calibrated since there are two PCRS devices on opposite sides of SST’s field of view. This process cannot control for high- or mid-frequency disturbances, but it is effective in compensating for slow thermal drifts between the spacecraft and telescope. The system is extremely low power, since it uses only two 4×4 pixel arrays (Mainzer and Young 2004).

7.1 Pointing and Control Systems for Future Missions

For future missions with larger primary mirror diameters, achieving the required pointing stability may be challenging with present technology. Proposed pointing and control systems include fine guidance sensors integrated into the instrument packages of the observatory and a fine-steering mirror that allows for active adjustment of the observatory’s line of sight. Some of the proposed steering mirrors have reaction times fast enough that they can keep the observatory’s line of sight stable despite the presence of high-frequency jitter disturbances.



■ Fig. 9-38

A schematic of the pointing control system for JWST. Information from the fine guider system controls the FSM. Reaction wheels adjust the observatory pointing when necessary to keep the FSM near the center of its range (After Balzano et al. 2008)

Current missions have built structural stability into the interface between the spacecraft and the telescope, so that the offset between the star tracker and telescope lines of sight are fairly stable. (This ensures that guide stars fall within the fine guider field after a slew and enables star tracker data to be used to accurately control the “roll” position of the OTA.) As space telescopes become larger and more lightweight, it will be more challenging to achieve such structural stability. For the proposed 10-m ATLAST system, the telescope is connected to the spacecraft through a single long arm. The system requires an active measurement of the offset between the coarse-pointing system on the spacecraft and the fine guidance sensors in the telescope but has the advantage that the telescope is more isolated from mechanical disturbances on the spacecraft (Postman 2009).

If an observatory’s solar panels were mounted on adjustable booms, they could potentially be used as solar sails to nullify solar torques, or as solar rudders to dump momentum. This would eliminate the momentum-dumping process and increase observing efficiency. The proposed 8-m ATLAST mission has determined that a 10-m solar array boom extending from the spacecraft could allow for over 6 days of continuous high-precision pointing observation (Postman 2009).

8 Thermal Systems

Every component on a space telescope has been designed to work within a narrow range of operating temperatures. The thermal design of the telescope must ensure that every component is within its operating range despite the presence of external sources of heat (such as the Sun and the Earth) and internal sources of heat (such as the spacecraft electronics, moving parts, and sunshields).

For UV/optical observatories, most components can operate near room temperature with the exception of the instrument detectors, which must be cooler in order to minimize detector

noise. HST operates at a temperature of about 290 K (20°C), but the instrument detectors are cooled to lower temperatures to suppress detector noise sources. Infrared observatories must operate at colder temperatures so that the mirrors and structures within the telescope do not emit infrared radiation that falls on the detectors in the system. JWST will operate at 40 K (−230°C) to reduce the background out to wavelengths of 28 μm and SST operated at about 6 K (−270°C) to reduce the background out to wavelengths of 180 μm.

The operating temperatures must also be stable over time. Temperature stability prevents changes in the structure that could cause optics to misalign or distort; submicron changes in a mirror position can degrade the performance of the telescope. Structural changes can also compromise calibrations, such as the measured offset between the star trackers and the instrument line of sight. The observatory structures are designed to be mechanically stable and are athermalized as much as possible, but those processes are imperfect and cannot completely eliminate the need for temperature stability.

Thermal stability is a difficult requirement for an observatory in LEO, since it travels in and out of the shadow of the Earth approximately every 45 min. HST encounters a solar flux of approximately 1,300 W/m² when it is in the Sun, which drops to zero when HST is in the shadow of the Earth. Temperatures on the exterior of HST can vary from 190 to 330 K. In addition, the Earth reflects some of the Sun's thermal energy onto the telescope, and the warm Earth itself emits thermal radiation. JWST and SST experience a more stable thermal environment; because they are in solar orbits, the Sun is always in view and on the same side of the observatory. Thermal effects and reflection of sunlight from the Earth are also much less for telescopes in solar orbits.

8.1 Basic Components of Thermal Control Systems

Thermal control methods can either be active or passive. Active means of thermal control include systems that add or subtract heat from a component on command such as heaters, thermoelectric coolers (TECs), coolants, and cryocoolers. Missions that rely on cryogenics have lifetimes limited by the amount of coolant on board. Thermoelectric coolers (TECs) do not have lifetime limits but cannot achieve the extremely low temperatures needed for infrared missions. Mechanical cryocoolers also avoid lifetime limits but produce a great deal of jitter and require mechanical isolation from the rest of the system. (To date, cryocoolers have been used to cool individual instruments but have not been used to cool entire missions. The proposed SPICA mission would be the first (Sato et al. 2010).) Heaters can be used to add heat in order to control thermal gradients and to stabilize final operating temperatures, but the system must first be cooled to a temperature lower than the operational temperature by some other means (Donabedian 2003).

Passive means of thermal control are designed directly into the system and cannot be cycled on and off. The most obvious parts of a passive thermal control system are the outer shroud or sunshield (which protects the observatory from solar radiation) and radiators (large panels that radiatively transmit heat into space). Multilayer insulation blankets, which consist of layers of metalized plastic sheets, are used throughout the telescope to protect components from solar radiation and from thermal radiation from other components. The passive thermal control system also includes all materials, component shapes, joint types, and thermal coatings chosen for the observatory (Gilmore 2002).

8.2 Radiative and Conductive Heat Transfer

Radiative and conductive heat transfers determine the steady-state temperature of a space-based observatory. Materials and their shapes can be carefully chosen to control the conduction of heat through the telescope structure, and radiative emitters and insulators can be used to manipulate radiative transfer.

The energy transferred conductively through a plate of area A and thickness l can be written as

$$q_c = \frac{kA}{l} \Delta T, \quad (9.13)$$

where ΔT is the temperature difference between the front and back of the plate and k is the thermal conductivity of the material. Materials with low conductivities, small cross-sectional areas, and long lengths can be used to isolate parts of the system from conductive heat transfer. Conversely, materials with high conductivities can be used to “dump” heat from the telescope structure to its radiators or to transfer heat to coolants. The shapes of the materials can be chosen to further enhance or inhibit the heat transfer.

Radiative heat transfer involves absorption and emission of electromagnetic radiation. Materials tend to absorb primarily in short wavelengths and emit that heat in infrared wavelengths. For any material, α_s is the solar absorptance and gives the percentage of incident sunlight absorbed by the material. The infrared emissivity of the material, ϵ_{IR} , is the percentage of that energy that is radiated away in the infrared. It is generally accurate to assume that for infrared wavelengths, infrared emissivity is equal to the infrared absorption ($\epsilon_{IR} = \alpha_{IR}$). Both absorptance and emittance depend on temperature. The final temperature of a surface of area A depends on the total energy input (due to solar flux, infrared flux, and conductive heat, $A\alpha_s\Phi_s + A\alpha_{IR}\Phi_{IR} + q_c$) and the total energy radiated away from the surface ($A\epsilon_{IR}\sigma T^4$).

Observatory materials that are in the sunlight must not heat up and radiate onto the rest of the observatory. To accomplish this, materials with low solar absorption and high-infrared emissivity are used; the ratio of α/ϵ is low. Most sunlight is reflected by these materials. The little that is absorbed will be reemitted as infrared energy. Examples of materials with low α/ϵ include dielectric films on polished metals, white paints, and second-surface mirrors.

Radiators are designed to radiate heat from the observatory out into deep space. The heat dissipated by a radiator can be written as

$$q_r = \epsilon A \sigma T^4, \quad (9.14)$$

where A is the area of the radiator, T is the temperature of the radiator, and σ is the Stephan-Boltzmann constant ($5.67 \cdot 10^{-8} \text{ W}/(\text{m}^2 \text{ K}^4)$). To transfer heat efficiently, materials with high-infrared emissivity are chosen. At cryogenic temperatures, very large radiator areas are needed to dissipate even small amounts of power (Rohsenow et al. 1998). Often, radiators consist of coated metallic honeycomb structures.

To prevent radiative heat transfer to an object inside the observatory shroud, materials with low α and ϵ are used (radiative insulators). MLI accomplishes this by using low α/ϵ materials in the outer layer, followed by layers with low emissivity that prevent infrared radiation from reaching the objects behind the MLI. Thin polymer meshes are used between layers to prevent conductive heat transfer.

8.3 HST Thermal Design

Since HST observes in the ultraviolet and visible, the structure of the observatory can operate at room temperature without generating significant background in those wavelength regimes. To reach a stable temperature of about 290 K (20°C), most components in HST are passively cooled to below room temperature utilizing MLI blankets, thermal control coatings, and design to control conductive heat transfer. Heaters are then used to raise each component's temperature to the desired level (Goddard Space Flight Center 1993).

The primary mirror carries heaters to stabilize the temperature of the mirror at about 289 K (16°C). The central baffle of the system is colder (267 K or -6°C) and could cause radial thermal gradients in the primary mirror that could distort the shape of the mirror. MLI layers and heaters at the inner diameter of the primary mirror prevent such gradients. MLI blankets also protect the primary from the aft shroud temperature variations.

The secondary mirror is also actively heated. The main baffle, central baffle, and secondary baffles (shown in [Fig. 9-12](#)) are all passively cooled. All baffles facing the secondary use low-emittance materials or MLI to prevent them from thermally radiating onto the secondary.

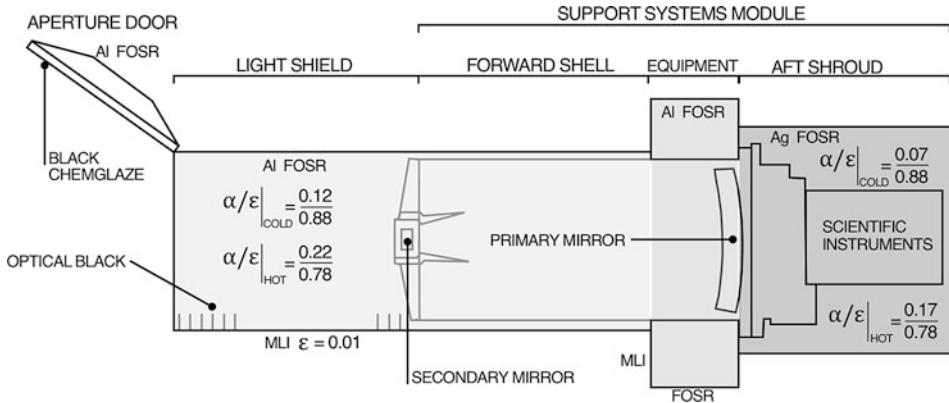
A graphite-epoxy truss called the "Metering Truss" holds the secondary and primary mirrors in alignment and is carefully designed to minimize bending in the structure with temperature. The temperature of the truss is passively controlled; each leg is enclosed in MLI. (The truss length has changed slowly over time following an exponential decay pattern that suggests an outgassing process. The secondary mirror has been adjusted to compensate (Lallo 2012).)

On the side of HST that carries the aperture door (which is usually in direct sunlight), the external surfaces are covered with MLI blankets which have an outer layer of aluminum or silver Teflon Flexible Optical Solar Reflector (FOSR), followed by many layers of aluminized Kapton, and an inner layer of a dark material for stray light control, as sketched in [Fig. 9-39](#). The side of the observatory that is opposite the aperture door is usually not in direct sunlight, but the infrared load from the Earth passes across it periodically. That side of the external shell is covered with MLI (without the FOSR) with an emittance of 0.01 so that it does not absorb the Earth's IR load.

Heaters are used to bring the instruments up to room temperature from the cooler temperature of the aft shroud. Most of the instrument's detectors, though, are cooled with thermoelectric coolers to suppress detector noise sources. The detectors in the NICMOS instrument are cooled with a cryocooler, and the ACS, STIS, and WFC3 CCDs are cooled by thermoelectric coolers. A 6-stage TEC stack, shunted to a radiator via a set of single-stage TECs run in parallel, cools the WFC3 IR detector to 145 K.

Problematically, the aft shroud on HST was found to be warming at a rate of about 1.2 K per year. Engineers speculated that this was due to degradation of the aluminum FOSR in the orbital environment. If temperatures in the shroud rise above the operating temperatures of the instruments, thermal control of the instruments can be lost (Hopkins et al. 2002). An aft shroud cooling system was installed in the third servicing mission in order to correct the problem.

Solar arrays generate large amounts of heat, and for missions in LEO, they undergo rapid temperature changes due to the day-night cycle. To combat this, HST's arrays are connected to the forward shell with rods of insulating materials (see [Fig. 9-11](#)). The aluminized Teflon FOSR that encloses the forward shroud protects against radiative transfer from the solar arrays to the telescope. The first generation of solar arrays on HST produced large amounts of mechanical jitter caused by thermal stresses induced by the day-to-night temperature swings. The jitter typically caused the FGSS to lose lock on the guide stars, which caused an interruption to



■ Fig. 9-39

Materials in the outer enclosure of the Hubble Space Telescope are designed to control the temperature of the telescope as it passes in and out of view of the Sun and as the thermal load from the Earth rotates about the outer shell. The aperture door, black paints, and baffles are present to help control stray light (After Lockheed Missiles and Space Company 1985)

the science exposures, resulting in degraded data quality and observatory efficiency. A second generation of stiffer arrays was installed to correct the problem.

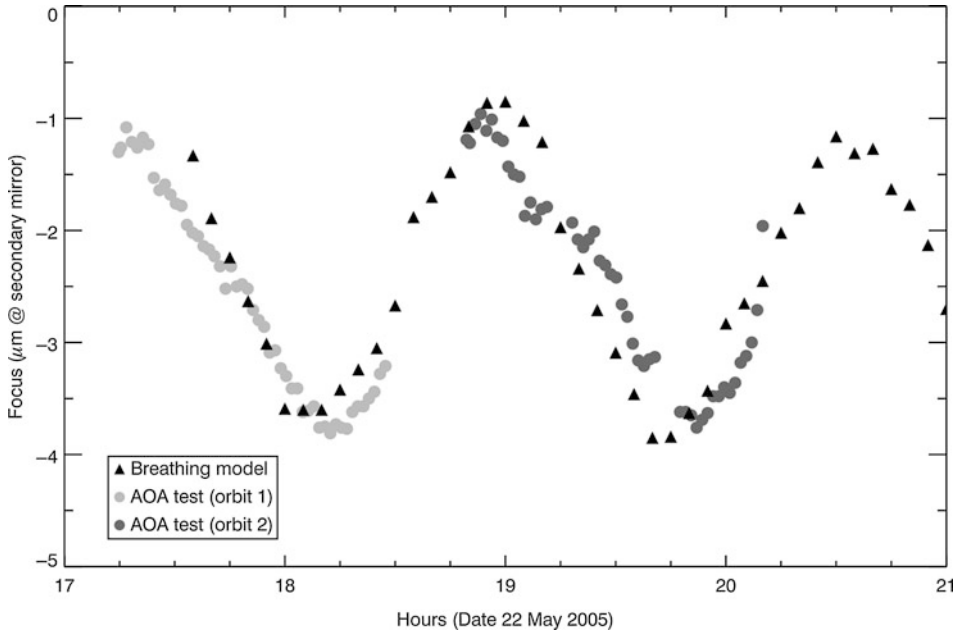
The focus of HST's telescope varies with time, and an example is shown below in [Fig. 9-40](#) (Lallo 2012). The data correlate with temperature changes in the vicinity of the secondary mirror support structure caused by radiative loads. Temperature changes in the telescope are largely driven by moving in and out of the shadow of the Earth, changes in the angle of the Sun with respect to the telescope, and the infrared load from the Earth when the telescope pointing is such that it looks directly at the Earth for part of every orbit.

8.4 SST Thermal Design

SST pioneered the use of passive cooling for infrared space telescopes. SST was initially envisioned as an actively cooled telescope in low Earth orbit. After cost constraints forced a descoping of the mission, the new, smaller launch vehicle meant that the mass and volume of the mission had to be cut by more than 50%. The resulting thermal design was driven by several goals:

- Reduce the cryogen amount in order to fit into the mass and volume dictated by the launch vehicle
- Cool the instruments sufficiently to operate in a background-limited condition out to wavelengths of $180 \mu\text{m}$
- Provide adequate cryogen to achieve a 2.5-year mission lifetime

To accomplish these challenging goals, the basic thermal design was changed to include a passively cooled telescope attached to a cryogen tank that actively cooled the instruments to 1.25 K. The vented superfluid helium was used to further cool the telescope to its operational temperature of 4–10 K once it was in orbit. The design was revolutionary; previous infrared missions actively cooled both the telescope and the instruments (Davies 2006). The change to a



■ Fig. 9-40

A typical change in focus over an HST orbit is seen here to be $\sim 3 \mu\text{m}$ of secondary mirror motion or $\sim 18 \text{ nm}$ RMS wavefront error. Round gray points are measurements of a single PSF in individual exposures taken with the Advanced Camera for Surveys' High-Resolution Channel. The black triangular points are values from an empirical model (Lallo 2012)

passively cooled telescope allowed the telescope to launch while warm and drastically reduced the amount of cryogen needed for the mission since only the instruments were actively cooled.

A schematic of the SST observatory is shown in ► Fig. 9-14. The “Outer Shell” section of the observatory is passively cooled to 40 K. On the space-facing side of the outer shell, a high-emissivity material is used to radiate heat into deep space. On the Sun-facing side of the outer shell, the solar panel blocks direct solar flux onto the shell and telescope, and the outer shell is polished aluminum. The choice of a drift-away orbit eliminates the Earth as a source of emitted thermal and reflected solar radiation. A radiatively cooled panel with low emissivity (labeled “Solar Panel Shield” in the figure) protects the outer shell from the warm solar panel (Lee et al. 1998).

The CTA (telescope, instrument, and cryostat assembly) is mechanically mounted to the spacecraft using low-conductivity struts for thermal isolation. A radiatively cooled panel shields the outer shell of the CTA from the spacecraft, which operates at room temperature.

The instrument chamber is mounted to the 360-L helium tank, which provides a 1.4-K heat sink for the science instruments. The instruments comprise the majority of the 6-mW heat load on the cryogen. The cryogen determines the science mission lifetime, so everything possible was done to prevent unintended heat loads on the cryogen. The tank is enclosed in two vapor-cooled shields, 45 layers of MLI, and an external guard vacuum shell. The helium gas vented from the tank is routed along the cryostat's shields, the vacuum shell, and the outer shell of the

CTA. This additional cooling brings the outer shell and the Telescope temperatures down to the final operational temperature of 4–10 K. It also prevents the outer shell from placing additional heat load on the cryogen.

8.5 JWST Thermal Design

JWST is designed to be passively cooled to approximately 40 K, including the instrument detectors, primarily because both the telescope and instruments were physically too large to be cooled with cryogenics or cryocoolers. Since no cryogenics will be onboard, the maximum lifetime of JWST is set by the propellant used to conduct orbital maintenance maneuvers. The fully passive design requires highly accurate thermal models; if a cryogen-cooled system runs warmer than planned, the lifetime of the mission is shortened, but if a passively cooled system runs warmer than planned, the entire mission could be compromised due to increased detector noise.

JWST's thermal design relies on a multilayer sunshield to separate the warm components in the observatory (the spacecraft and the solar array) from the cold components (the telescope and instruments) as sketched in [Fig. 9-41](#). The telescope and the instruments are always in the shade of the sunshield and are not enclosed in an outer shroud. The sunshield reduces 200 kW of incident solar radiation to milliWatts of radiation incident on the telescope and instruments (Gardner et al. 2006). The outer layers use a material with a low α/ϵ , and inner layers provide a low- ϵ material to prevent radiative heat transfer.

On the warm side of the observatory, much of the spacecraft bus is covered with MLI. Radiator panels keep the bus from overheating, and shades over the panels keep Sunlight away from the radiators.

On the cold side of the observatory, a deployed tower moves the telescope and instruments away from the warm side of the observatory. This is mainly so that the primary and secondary

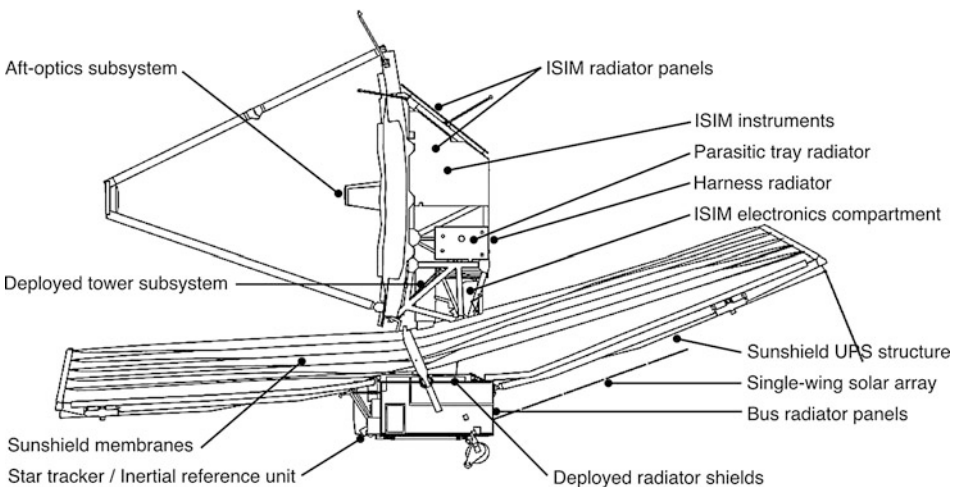


Fig. 9-41

A sketch of JWST showing relevant thermal control components (NASA 2009, labels added)

mirrors “see” less of the sunshield. The primary, secondary, and secondary struts all cool passively. The tertiary and FSM in the Aft Optics Subsystem are cooled using radiators mounted onto the structure.

The instruments and detectors in the ISIM compartment produce about 300 mW of heat, which is radiated away using more than 8 m² of radiator panels located behind the primary mirror (Parrish 2004). The MIRI instrument, which will observe in wavelengths out to 28 μm, is cooled to its operational temperature of 5–7 K using a cryocooler. It is the only instrument that is actively cooled.

A warm electronics box is also located on the “cold” side of the observatory, underneath the instrument compartment. It operates at 300 K and has dedicated radiators and baffles to prevent the heat from radiating onto the cold components.

8.6 The Future in Thermal Designs

Cryogenics severely limit mission lifetime and are a significant fraction of the mass and volume for any mission. Future NIR missions will likely continue to avoid the use of cryogenics. It is difficult to passively cool to temperatures below about 40 K, though, so far-infrared missions will continue to require active cooling systems.

Fully passive cooling strains current thermal design and modeling capabilities. As observatories get larger, it is increasingly difficult to create thermal chambers that can hold the full observatory and recreate on orbit conditions, so ground testing of on orbit thermal environments will not be possible. To accurately predict on orbit temperatures and background levels, future passively cooled systems will require improvements in thermal modeling capabilities, as well as extremely careful thermal designs in which every interface in the system is designed to be both functional and have known thermal properties. Future observatories could be designed to compensate for structural changes due to thermal uncertainties through the more extensive use of continuous active and adaptive control of the optical alignment as described in [▶ Table 9-10](#). With a fully active and adjustable telescope, ground testing could be limited to validation (through detailed modeling or the testing of smaller subsystems) that the telescope can be assembled or deployed to within the capture range of the onboard alignment system.

9 Conclusion

Space telescopes have made, and will continue to make, fundamental discoveries about the physical nature of the cosmos. Space telescopes can produce diffraction-limited images of very faint objects at any wavelength and over wide fields of view. They are free from the atmospheric extinction, high background levels, and the atmospheric turbulence that limit ground-based observations. Although space telescopes are currently limited to smaller aperture sizes, they can still achieve higher spatial resolutions and sensitivities than their ground-based counterparts.

HST, SST, and other current space telescopes have demonstrated that the ability to hold a large optical system on target and in focus for extended exposure times is now readily achievable from free-flying space observatories. JWST is the next step in the evolution of space telescope

technology – a large segmented primary mirror with integrated wavefront sensing and control systems.

The search for biosignature around nearby stars and the future demands of observational astrophysics will drive the need to find ways to cost-effectively build, launch, and operate optical and infrared telescopes that are even larger than JWST. To realize future large space telescopes such as the 10-m ATLAST mission will take all the approaches described and push the technologies and techniques one (or two) steps further. To keep the launch mass and volume within reasonable bounds, lightweight optics and structures will be required, which in turn will require high-bandwidth active mirror control to carry out on orbit alignment and possibly high-bandwidth adaptive control of the optical surfaces to reach the diffraction limit in optical and UV wavelengths. Integrating and testing such large telescopes becomes increasingly difficult as the limits of ground test equipment and thermal chambers that can recreate on orbit conditions are reached. More sophisticated designs and models (optical, mechanical, and thermal) of the entire observatory system will be required. In addition, higher data storage and downlink bandwidths will be needed for the larger detector arrays of future missions. Advances in heavy lift launch capability will need to be monitored and factored into observatory design as well.

While costly, the scientific impact of space observatories is substantial. Space telescopes working in conjunction with major ground-based telescopes have revolutionized our understanding of star and galaxy formation, of the nature of the universe, and have begun to address the question of whether there are other Earth-like planets around other stars. The future generation of space observatories will extend this scientific inquiry to the earliest epoch when the very first stars formed and will, ultimately, allow us to answer the question “Is there life elsewhere in the Galaxy?”

Acknowledgments

We would like to thank many people that contributed greatly to this effort. Jill Lagerstrom provided invaluable help with the research, reference formats, and measurements of observatory papers and citations. Amy Gonigam and Elizabeth Fraser in the STScI library found every book and article that we could possibly have needed. George Hartig and Matt Lallo are reservoirs of HST program, hardware, and operation knowledge and also reviewed the chapter. Carl Biagetti kindly reviewed the chapter, and Remi Soummer reviewed the high-contrast imaging section. Chris Long and Tom Wheeler shared their highly detailed knowledge of HST’s hardware. Roeland van der Marel wisely directed us to Bely’s excellent book. Ed Nelan answered many pointing and tracking questions and expertly reviewed the section for us. Scott Friedman answered several general astronomy questions. Elizabeth Barker helped find reasonable assumptions for detector noise levels. Carol Christian provided current information on the Davidson metric. Tracy Beck and Marianne Takamiya (University of Hawai’i Hilo) provided information on Gemini’s seeing and background. Erick Young (SOFIA Science Center, Ames Research Center) graciously answered SST questions and reviewed the chapter. Conrad Schiff (GSFC) answered orbital questions. Randy Frank (BATC) was a valuable resource on thermal designs. Paul Lightsey (BATC) chatted with us about design choices for both HST and JWST. Chuck Bowers (GSFC) helpfully tracked down several JWST details for us. David Content (GSFC) and Robert Egerman (ITT) provided information on corrugated mirror technology.

References

- Apai, D., Lagerstrom, J. P., Reid, I. N., et al. 2010, *PASP*, 122, 808
- Bahcall, J. N., & Soneira, R. M. 1984, *ApJS*, 55, 67
- Balzano, V., Isaacs, J. C., & Nelan, E. P. 2008, *Proc. SPIE*, 7016, 70161X
- Bate, R. R., Mueller, D. D., & White, J. E. 1971, *Fundamentals of Astrodynamics* (New York: Dover)
- Bely, P. Y. 2003, *The Design and Construction of Large Optical Telescopes* (New York: Springer)
- Casani, J., et al. 2010a, JWST Test Assessment Team (TAT) Final Report. http://www.jwst.nasa.gov/resources/JWST_TAT_Final_Report_100907.pdf
- Casani, J., et al. 2010b, JWST Independent Comprehensive Review Panel (ICRP) Final Report. http://www.nasa.gov/pdf/499224main_JWST-ICRP_Report-FINAL.pdf
- Cash, W., Kendrick, S., Noecker, C., et al. 2009, *Proc. SPIE*, 7436, 743606
- Chaney, D. M., Hadaway, J. B., Lewis, J., Gallagher, B., & Brown, B. 2011, *Proc. SPIE*, 8150, 815008
- Christian, C. A., & Davidson, G. 2006, in *ASSL 335, Organizations and Strategies in Astronomy* Vol. 6, ed. A. Heck (New York: Springer), 145
- Clampin, M. 2011, Overview of the James Webb Space Telescope Observatory. *Proc. SPIE*, 8146, 814605
- Davies, J. K. 1997, *Astronomy from Space: The Design and Operation of Orbiting Observatories* (New York: Wiley)
- Davies, J. K. 2006, *The Space Review*. <http://www.thespacereview.com/article/688/1>
- Den Hartog, J. P. 1961, *Mechanics* (New York: Dover)
- Donabedian, M. 2003, *Spacecraft Thermal Control Handbook, Volume 2: Cryogenics* (El Segundo: Aerospace Press)
- Doxsey, R. E. 2006, *Proc. SPIE*, 6270, 627006
- Egerman, R., De Smitt, S., & Strafford, D., 2010, Low-weight, low-cost, low-cycle time replicated glass mirrors. *Proc. SPIE*, 7739, 77390G
- Elachi, C., 1987, *Introduction to Physics and Techniques of Remote Sensing* (New York: Wiley)
- Gallagher, D. B., Irace, W. R., & Werner, M. W. 2003, *Proc. SPIE*, 4850, 17
- Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, *SpaceSciRev*, 123, 485
- Gerchberg, R. W., & Saxton, W. O. 1972, *Optik*, 35, 237
- Gehrz, R. D., Roellig, T. L., Werner, M. W., et al. 2007, *Rev. Sci. Instrum.*, 78, 011302
- Giacconi, R. 2008, *Secrets of the Hoary Deep* (Baltimore: Johns Hopkins University Press)
- Gillet, F. C., & Mountain, M. 1998, in *ASP Conf. Ser. 133, Science with the NGST (Next Generation Space Telescope)*, eds. E. P. Smith & A. Koratkur (San Francisco: ASP), 42
- Gilmore, D. G. 2002, *Spacecraft Thermal Control Handbook, Volume 1: Fundamental Technologies* (El Segundo: Aerospace Press)
- Goddard Space Flight Center 1993, *OTA/FGS Thermal Control System Description and Operating Manual* (Greenbelt: NASA GSFC)
- Goddard Space Flight Center 2010, *On-Orbit Satellite Servicing Study: Project Report* (Greenbelt: NASA GSFC)
- Guyon, O. 2005, *ApJ*, 629, 592
- Hickey, G., Barbee, T., Ealey, M., & Redding, D. 2010, *Proc. SPIE*, 7731, 773120
- Hopkins, R. A., Schweickart, R. B., Finley, P. T. & Volz, S. M. 2002, *Proc. SPIE*, 4850, 42
- Kasdin, N. J., Cady, E. J., Dumont, P. J., et al. 2009, *Proc. SPIE*, 7440, 744005
- Kasting, J., Traub, W., Roberge, A., et al. 2009, *arXiv:0911.2936v1*
- Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, *ApJ*, 713, L79
- Krist, J. E., & Burrows, C. J. 1995, *Appl. Opt.*, 34, 4951
- Lallo, M. 2012, Experience with the Hubble Space Telescope: 20 years of an archetype. *Opt. Eng.*, 51, 011011
- Lawson, P. R., Traub, W. A., & Unwin, S. C. 2009, *Exoplanet Community Report*, JPL Publication 09-3 (Pasadena: Jet Propulsion Laboratory)
- Lee, J. H., Blalock, W., Brown, R. J., Volz, S. M., Yarnell, T., & Hopkins, R. A. 1998, *Proc. SPIE*, 3435, 172
- Lester, D. F., Benford, D. J., Blain, A., et al. 2004, *Proc. SPIE*, 5487, 1507
- Lightsey, P., Atkinson, C., Clampin, M., & Feinberg, L. 2012, James Webb Space Telescope: large deployable cryogenic telescope in space. *Opt. Eng.*, 51, 011003
- Lo, A. S., Glassman, R., Dailey, D., et al. 2010, *Proc. SPIE*, 7731, 77312E
- Lockheed Missiles and Space Company, 1985, *Space Telescope Systems Description Handbook, ST/SE-02* (Sunnyvale: Lockheed Missiles & Space Company)
- Mainzer, A. K., & Young, E. T. 2004, *Proc. SPIE*, 5487, 93
- Matthews, G., Barrett, D., Bolton, J., et al. 2003, *Proc. SPIE*, 5180, 169
- Mountain, M., van der Marel, R., Soummer, R. et al. 2009, *arXiv:0909.4503v1*
- Nakagawa, T. 2008, *Proc. SPIE*, 7010, 70100H

- Oke, J. B. 1974, *ApJS*, 27, 21
- Oswald, H. W., Siegmund, B. Y., Welsh, C. M., et al. 2004, *Proc. SPIE*, 5488, 13
- Parrish, K. 2004, in *From Spitzer to Herschel and Beyond: The Future of Far-Infrared Space Astrophysics Conference* (Pasadena). <http://safir.jpl.nasa.gov/BeyondSpitzerConf/index.shtml>
- Pilbratt, G. L. 2008, *Proc. SPIE*, 7010
- Postman, M. 2009, arXiv:0904.0941v2
- Prussing, J. E., & Conaway, B. A. 1993, *Orbital Mechanics* (New York: Oxford University Press)
- Ramsay, S. K., Mountain, C. M., & Geballe, T. R. 1992, *MNRAS*, 259, 751
- Rohsenow, W. M., Hartnett, J. P., & Cho, Y. I. 1998, *Handbook of Heat Transfer* (3rd ed.; New York: McGraw-Hill)
- Sabelhaus, P. A., Campbell, D., Clampin, M., et al. 2005, *Proc. SPIE*, 5899, 58990P
- Sato, et al. 2010, Conceptual design of a cryogenic system for the next-generation infrared space telescope SPICA. *Proc. SPIE*, 7731
- Schroeder, D. J. 1987, *Astronomical Optics* (San Diego: Academic)
- Space Telescope Science Institute 2010, *FGS Instrument Handbook for Cycle 19* (Baltimore: STScI). <http://www.stsci.edu/hst/fgs/design/hst/fgs/documents/instrumenthandbook/>
- Space Telescope Science Institute 2011, *Hubble Space Telescope Primer for Cycle 19* (Baltimore: STScI). <http://www.stsci.edu/hst/proposing/documents/primer/>
- Spitzer Science Center, 2011, *Spitzer Space Telescope Handbook, Version 2.0* (Pasadena: SSC). <http://irsa.ipac.caltech.edu/data/SPITZER/docs/spitzermission/missionoverview/spitzertelescopehandbook/home/>
- Stahl, H. P. 2010, *Opt. Eng.*, 49, 053005
- Strafford, D. N., DeSmitt, S. M., Kupinski, P. T., & Sebring, T. A. 2006, *Proc. SPIE*, 6273, 62730R
- Thomson, W. T. 1986, *Introduction to Space Dynamics* (New York: Dover)
- Thronson, H., Lester, D., Watson, J., & Moe, R. 2005, in *Report of the Space Resources Roundtable VII* (Houston, TX: Lunar and Planetary Institute), 104
- Trauger, J. T., & Traub, W. A. 2007, *Nature*, 446, 771

10 CMB Telescopes and Optical Systems

*Shaul Hanany*¹ · *Michael D. Niemack*^{2,3} · *Lyman Page*⁴

¹School of Physics and Astronomy, University of Minnesota, Minneapolis, MN, USA

²National Institute of Standards and Technology, University of Colorado, Boulder, CO, USA

³Physics Department, Cornell University, Ithaca, NY, USA

⁴Department of Physics, Princeton University, Princeton, NJ, USA

1	<i>Introduction</i>	432
1.1	Celestial Emission at CMB Frequencies	433
1.2	Instrument Resolution	434
1.3	Throughput and Modes	437
1.4	Noise	439
1.5	Polarization Terminology	440
2	<i>Ground- and Balloon-Based Systems</i>	442
2.1	General Considerations	443
2.2	Balloon- Versus Ground-Based Systems	448
2.3	Arrays of Detectors and Increases in DLFOV	450
2.4	Refractor-Based Optical Systems	452
2.5	Reflector-Based Optical Systems	453
2.6	Polarization Properties	455
3	<i>Large Ground-Based Telescopes, ACT and SPT</i>	459
3.1	Detailed Optical System Comparison	459
3.2	ACTPol and SPTPol	465
4	<i>Interferometers</i>	465
5	<i>The CMB Satellites</i>	467
5.1	Relikt	469
5.2	COBE	470
5.3	WMAP	472
5.4	Planck	474
5.5	A Future Satellite	476
	<i>Acknowledgments</i>	476
	<i>References</i>	476

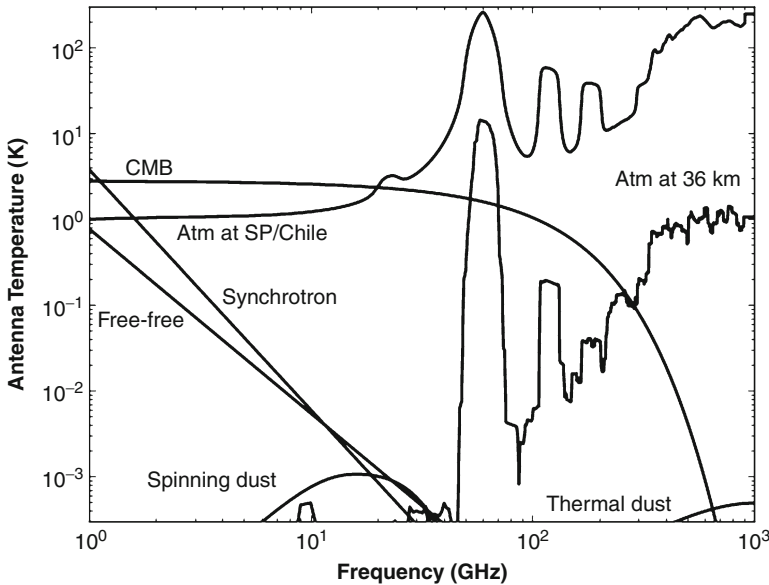
Abstract: The cosmic microwave background radiation (CMB) is now firmly established as a fundamental and essential probe of the geometry, constituents, and birth of the observable universe. The CMB is a potent observable because it can be measured with precision and accuracy. Just as importantly, theoretical models of the universe can predict the characteristics of the CMB to high accuracy, and those predictions can be directly compared to observations. There are multiple aspects associated with making a precise measurement. In this chapter, we focus on optical components for the instrumentation used to measure the CMB polarization and temperature anisotropy. We begin with an overview of general considerations for CMB observations and discuss common concepts used in the community. We next consider a variety of alternatives available for a designer of a CMB telescope. Our discussion is guided by the ground- and balloon-based instruments that have been implemented over the years. In the same vein, we compare the arc-minute resolution Atacama Cosmology Telescope (ACT) and the South Pole Telescope (SPT). CMB interferometers are presented briefly. We conclude with a comparison of the four CMB satellites, Relikt, COBE, WMAP, and Planck, to demonstrate a remarkable evolution in design, sensitivity, resolution, and complexity over the past 30 years.

Keywords: CMB, Telescopes

1 Introduction

The tremendous scientific payoff from studies of the cosmic microwave background radiation (CMB) has driven researchers to develop new detectors and new detection techniques. For the most part, CMB measurements have been made with dedicated instruments in which the optical elements are designed specifically to mate with the detectors (rather than in facility-type telescopes). The instruments run the gamut of radio- and mm-wave detection techniques: heterodyne receivers, direct power receivers, correlation receivers, interferometers, Fourier transform spectrometers, and single- and multimode bolometric receivers. The quest for ever more sensitive measurements of the CMB, including its polarization, has led to the development of arrays of hundreds to thousands of detectors, some of which are polarization sensitive. These arrays are coupled to unique, large-throughput optical systems. In this article, we will focus primarily on optical systems for instruments that are used to measure the temperature anisotropy and polarization of the CMB. In other words, instruments that are designed to measure only the temperature difference or polarization as a function of angle on the sky. We refer the reader to the chapter by Mather et al. (2012) in these volumes for more information about other aspects of the CMB.

Before explicitly discussing the optical systems, we introduce in this section the celestial emission spectrum at CMB frequencies, discuss how the instrument resolution is determined, and present the angular power spectrum. We then introduce the concepts of throughput and modes and end with a discussion of the limits imposed by system noise because it is one of the driving considerations for any optical design. In [▶ Sect. 2](#), we review the various choices available for a CMB optics designer and the main optical systems that have been used to date. We also discuss more recent developments with the introduction of large focal plane arrays and the efforts to characterize the polarization of the CMB. The ACT and SPT instruments are the highest resolution telescopes dedicated to CMB measurements to date. They are also good examples for the state-of-the-art in CMB optical design at the time of their design, mid-decade 2000. They are described and compared in [▶ Sect. 3](#). CMB interferometers are briefly presented in [▶ Sect. 4](#), and the optical systems of the four CMB satellites to date are reviewed in [▶ Sect. 5](#).



■ Fig. 10-1

Sources of sky emission between 1 and 1000 GHz for a region of sky near a galactic latitude of roughly 20° . The flat part of the CMB below ~ 30 GHz is called the Rayleigh-Jeans portion. A Rayleigh-Jeans source with frequency-independent emissivity would be indicated by a horizontal line on this plot. The synchrotron emission is from cosmic rays orbiting in galactic magnetic fields and is polarized. Free-free emission is due to “breaking radiation” from galactic electrons and is not polarized. The amplitude of the spinning dust is not well known. This particular spinning dust model comes from Ali-Haïmoud et al. (2009). The standard spinning dust emission is not appreciably polarized. Thermal emission from dust grains, which is more intense than the CMB above ~ 700 GHz, is partially polarized. The atmospheric models are based on the ATM code (Pardo et al. 2001); they use the US standard atmosphere and are for a zenith angle of 45° . The Atacama/South Pole spectrum is based on a precipitable water vapor of 0.5 mm. The difference between the two sites is inconsequential for this plot. The atmospheric spectra have been averaged over a 20% bandwidth. The pair of lines at 60 and 120 GHz are the oxygen doublet. The lines at 19 and 180 GHz are vibrational water lines. The finer scale features are from ozone

1.1 Celestial Emission at CMB Frequencies

► *Figure 10-1* shows the antenna temperature of the sky from 1 to 1,000 GHz for a region at a galactic latitude of roughly 20° . Ignoring emission from the atmosphere, synchrotron emission dominates celestial emission at the low frequency end and dust emission dominates at high frequencies. These galactic emission components may be different by an order of magnitude depending on galactic longitude. The CMB radiation dominates emission between about 20 and 500 GHz. The experimental challenge is, however, to measure spatial fluctuations in the CMB at parts in 10^6 or 10^7 of the level, a couple of orders of magnitude below the bottom of the plot. The polarization signals are lower than the temperature anisotropy by a factor of 10, and they too beckon to be measured to percent-level precision. The instrumental passbands, typically

20–30%, are chosen to avoid atmospheric emission lines or to help identify and subtract the foreground emission.

The basic picture in [Fig. 10-1](#) has remained the same for over 30 years (Weiss 1980), though over the past decade there has been increasing evidence for a new component of celestial emission in the 30 GHz region (e.g., Kogut et al. 1996; de Oliveira-Costa et al. 1997; Leitch et al. 1997). This new component is spatially correlated with dust emission and has been identified with emission by tiny grains of dust that are spun up to GHz rotation rates; hence, it has been dubbed “spinning dust.” A variety of mechanisms have been proposed for spinning up the grains (Draine and Lazarian 1998, 1999). Still, though, it is not clear that the source is predominantly spinning dust. Understanding this emission source is an active area of investigation.

1.2 Instrument Resolution

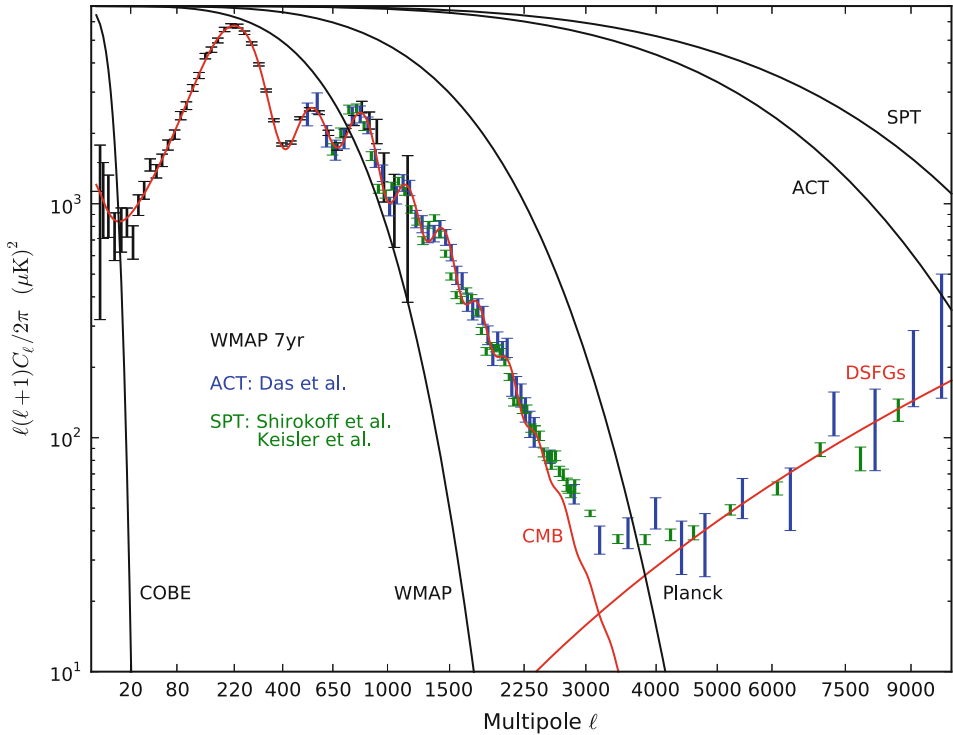
The resolution of a CMB telescope is easiest to think about in the time reversed sense. We imagine that a detector element emits radiation. The optical elements in the receiver direct that beam to the sky or onto the primary reflector. The size of the beam at the primary optic determines the resolution of the instrument. Such a primary optic can be a feedhorn that launches a beam to the sky, a lens, or a primary reflector. The connection between the spatial size of the beam at the primary optic and the resolution can be understood through the Fraunhofer’s diffraction relation (e.g., Born and Wolf 1980; Hecht 1987),

$$\psi(\theta) \propto \int_{\text{apt}} \psi_a(r) e^{kr \sin \theta \cos \phi} r dr d\phi, \quad (10.1)$$

where $\psi_a(r)$ is the scalar electric field (e.g., one component of the electric field) in the aperture or on the primary optic, $\psi(\theta)$ is the angular distribution of the scalar electric field in the far field ($d \gg 2D^2/\lambda$), and $k = 2\pi/\lambda$. The integral is over the primary reflector or, more generally, the aperture. For simplicity, we have taken the case of cylindrically symmetric illumination with coordinates r and ϕ for a circular aperture of diameter D , although a generalization is straightforward. The normalized beam profile is then given by $B(\theta) = |\psi(\theta)|^2/|\psi(0)|^2$. That is, if the telescope scanned over a point source very far away, the output of the detector as measured in power would have this profile as a function of scan angle θ . [Equation 10.1](#) gives an excellent and sometimes sufficient estimate of the far field beam profile.

To be more specific, let us assume that the aperture distribution has a Gaussian profile so that the integrals are simple. That is, $\psi_a(r) = \psi_0 e^{-r^2/2\sigma_r^2}$. We also assume that ψ_a is negligible at $r \geq D/2$ so that we can let the limit of integration go to infinity. This is the “large edge taper” limit. In reality, no aperture distribution can be Gaussian and some are quite far from it. The integral evaluates to $\psi_0 \sigma_r^2 e^{-\sigma_r^2 k^2 \sin^2(\theta)/2}$. For small angles, $\theta \sim \sin(\theta)$, and we find from the above that $B(\theta) = e^{-\theta^2/2\sigma_B^2}$ where $\sigma_B = \lambda/\sqrt{8}\pi\sigma_r$. In angular dimensions, the beam profile is most often characterized by a full width at half maximum, or twice the angle at which $B(\theta) = 1/2$. We denote this as $\theta_{1/2}$ and find $\theta_{1/2} = \sqrt{8 \ln(2)} \sigma_B = \sqrt{\ln(2)} \lambda/\sigma_r \pi$. We see the familiar relation that the beam width is proportional to the wavelength and inversely proportional to the size of the illumination pattern on the primary reflector with a prefactor that depends on the geometry. For this far-field Gaussian profile, the beam solid angle is $\Omega_B = \int B d\Omega = 2\pi\sigma_B^2$.

The natural “observable” for anisotropy measurements is the angular power spectrum for the following reason. When the distribution of the amplitudes of the fluctuations is Gaussian, as it apparently is for the primary CMB, *all* information about the sky is contained in the power



■ Fig. 10-2

Current best published measurements of the CMB temperature power spectrum (data points, Komatsu et al. 2011; Shirokoff et al. 2011; Das et al. 2011; Keisler et al. 2011) and a Λ CDM cosmological model (solid red, up to $\ell = 3,000$). The model power spectrum for $\ell > 3,000$ is due Poisson noise from confusion-limited dusty star-forming galaxies (DSFGs) at 150 GHz. The x-axis is scaled as $\ell^{0.45}$ to emphasize the middle part of the anisotropy spectrum. Gaussian approximations to the window functions are shown for COBE (7°), WMAP ($12'$), Planck ($5'$), ACT ($1.4'$, Swetz et al. 2011), and SPT ($1.1'$, Schaffer et al. 2011). The large size of the WMAP error bars near $\ell = 2$ and 1,000 are due to “cosmic variance” and finite beam resolution, respectively

spectrum. If there are correlations in the signal, for example, if the cooler areas had a larger spatial extent than the warmer areas or discrete sources of emission were clustered together, then higher-order statistics would be needed to fully describe the sky. Even in this case, the power spectrum is the best first-look analytic tool for assessing the sky. Searches for “non-Gaussianity” are an active area of research. While there are many possible sources of non-Gaussianity, the primary CMB anisotropy appears to be Gaussian to the limits of current measurements (e.g., Komatsu et al. 2011). A snapshot of the latest measurements of the power spectrum is shown in [Fig. 10-2](#).

The instrument resolution as expressed in the power spectrum is obtained from the Legendre transform of $B^2(\theta)$. To appreciate this, we take a step back and describe the connection between the observable, that is, the angular power spectrum and the antenna pattern of the instrument. Because the CMB covers the full sky, it is most usefully expressed as an expansion in spherical harmonics. The monopole term ($\ell = 0$) has been determined by COBE/FIRAS

to be $T_{\text{CMB}} = 2.725 \pm 0.001 \text{ K}$ (Fixsen and Mather 2002; plotted in **Fig. 10-1**). The dipole term ($\ell = 1$) is dominated by the peculiar velocity of the solar system with respect to the cosmic reference frame. As we are primarily concerned with cosmological fluctuations, we omit these terms from the expansion and we write the fluctuations as

$$\delta T(\theta, \phi) = \sum_{\ell \geq 2, -\ell \leq m \leq \ell} a_{\ell m} Y_{\ell}^m(\theta, \phi). \quad (10.2)$$

To the limits of measurement, the CMB fluctuations appear to be statistically isotropic (e.g., Basak et al. 2006): they are the same in all directions and thus have no preferred m dependence. The overall variance of the CMB fluctuations is then given by

$$\langle \delta T^2(\theta, \phi) \rangle = \sum_{\ell \geq 2} \frac{2\ell + 1}{4\pi} \langle |a_{\ell m}|^2 \rangle = \sum_{\ell \geq 2} \frac{2\ell + 1}{2\ell(\ell + 1)} \frac{\ell(\ell + 1)}{2\pi} C_{\ell} \equiv \sum_{\ell \geq 2} \frac{2\ell + 1}{2\ell(\ell + 1)} \mathcal{B}_{\ell}, \quad (10.3)$$

where the factor of $2\ell + 1$ comes from the sum of the m values, all of which have the same variance, and the 4π comes from averaging over the full sky. Generally C_{ℓ} is called the power spectrum, but in cosmology the term is just as frequently used for \mathcal{B}_{ℓ} . These quantities are the primary point of contact between theory and measurements. Cosmological models provide predictions for C_{ℓ} ; experiments measure temperatures on a patch of the sky and provide an estimate of C_{ℓ} . The quantity most often plotted is \mathcal{B}_{ℓ} .¹ It is the fluctuation power per logarithmic interval in ℓ . The x-axis of the power spectrum is the spherical harmonic index ℓ . As a rough approximation, $\ell \approx 180/\theta$ with θ in degrees.

The process of measuring the CMB with a beam of finite size acts as a convolution of the intrinsic signal (**Eq. 10.2**) with the beam function, $B(\theta)$. The finite resolution averages over some of the smaller angular scale fluctuations and thereby reduces the variance given in (**Eq. 10.3**). By Parseval's theorem, a convolution in one space corresponds to a multiplication in the Fourier transform space. In our case, because we are working on a sphere with symmetric beams, Legendre transforms, as opposed to Fourier transforms, are applicable. We may think of the square of the Legendre transform of $B(\theta)$, B_{ℓ}^2 , as filtering the power spectrum. (There is one power of B associated with one temperature map.) The transform of $B(\theta)$ is given by


$$B_{\ell} = 2\pi \int B(\theta) P_{\ell}(\cos \theta) d \cos(\theta) = B_0 e^{\ell(\ell+1/2)/2\sigma_b^2}, \quad (10.4)$$

where P_{ℓ} is a Legendre polynomial and B_0 is a normalization constant.

A Gaussian random field is fully described by the two-point correlation function, $C(\theta)$ (the Legendre transform of the power spectrum), which gives the average variance of two pixels separated by an angle. The variance given in (**Eq. 10.3**) is the angular correlation function evaluated for zero angular separation between pixels. The general relation, including the effects of measuring with a beam of finite resolution, is

$$C_{\text{meas}}(\theta) = \langle \delta T_{\text{meas}}(\theta_1, \phi_1) \delta T_{\text{meas}}(\theta_2, \phi_2) \rangle = \sum_{\ell \geq 2} \frac{2\ell + 1}{4\pi} C_{\ell} P_{\ell}(\theta) W_{\ell}. \quad (10.5)$$

¹The factor of $\ell(\ell + 1)/2\pi$ (Bond and Efstathiou 1984), as opposed to the possibly more natural $\ell(2\ell + 1)/4\pi$ (Peebles 1994), is derived from the observation that the cold dark matter model, without a cosmological constant, approaches $\ell(\ell + 1)$ at small ℓ for a scalar spectral index of unity. Needless to say, the model that gave rise to the now-standard convention does not describe nature. Another choice would be $(\ell + 1/2)^2$ because the wavevector $k \rightarrow \ell + 1/2$ at high ℓ . There is not a widely agreed upon letter for the plotted power spectrum. We use \mathcal{B} for both "bandpower" and J. R. Bond who devised the convention. The term bandpower refers to averaging the \mathcal{B}_{ℓ} over a band in ℓ .

Here, W_ℓ is the “window function.” In this expression, the angle θ goes between directions “1” and “2.”² The window function encodes the effects of the finite resolution. For a symmetric beam, $W_\ell = B_\ell^2$.  **Figure 10-2** shows approximations to the window functions, assuming Gaussian-shaped beams, for COBE, ($\theta_{1/2} = 7^\circ$), WMAP ($\theta_{1/2} = 12'$), Planck ($\theta_{1/2} = 5'$), ACT ($\theta_{1/2} = 1.4'$), and SPT ($\theta_{1/2} = 1.1'$). One immediately sees the relation between the resolution and how well one can determine the power spectrum. For example, COBE, which we discuss in more detail below, was limited to large angular scales (low ℓ) because of its relatively low angular resolution.


1.3 Throughput and Modes

One of the key characteristics for any optical system is the “throughput” or étendue (or “A-Omega”) of the system. It is a measure of the total amount of radiation that an optical system handles. Using Liouville’s theorem, which roughly states that the volume of phase space is conserved for a freely evolving system, one can show that $A\Omega$ is conserved for photons as long as there is no loss in the system. This means that at each plane of the system the integral of the product of areal (dA) and angular distribution ($d\Omega$) of the radiation is constant. For example, let’s assume that the effective angular distribution of a beam launched from a primary optic of radius r is a top hat in angular extent with an apex angle defined as $\theta_{1/2} = 2\theta_0$, analogous to the $\theta_{1/2}$ definition for Gaussian distributions above. Then one obtains a throughput of

$$A\Omega = 2\pi^2 r^2 (1 - \cos \theta_0) \simeq \frac{\pi^2 D^2 \theta_{1/2}^2}{16}, \quad (10.6)$$




where the approximation holds for $\sin \theta_0 \simeq \theta_0$. If the angular distribution is a Gaussian with width σ_B ($\theta_{1/2} = \sqrt{8 \ln 2} \sigma_B$), then

$$A\Omega \simeq 2\pi^2 r^2 \sigma_B^2 \simeq \frac{\pi^2 D^2 \theta_{1/2}^2}{16 \ln 2} \approx \frac{\pi^2 D^2 \theta_{1/2}^2}{11}. \quad (10.7)$$

Here, we also assume that σ_B is small, so that the integration over the angular pattern gives appreciable contributions only for $\sin \sigma_B \simeq \sigma_B$. If the primary reflector has an effective diameter of 1 m and the beam has $\theta_{1/2} = 0.15^\circ$, then the throughput is $0.04 \text{ cm}^2 \text{ sr}$ (assuming  10.6). Let’s say this radiation is focused down to a feed with an effective collecting area of 1 cm^2 . Conservation of throughput implies that now $\theta_{1/2} = 13^\circ$. In other words, as you squeeze down the area, the radiation has to go through by focussing or concentrating; the solid angle increases. While $A\Omega$ is conserved for any lossless optical system, it has a specific value for a system that supports only a “single mode” of propagating radiation:

$$A\Omega = \lambda^2, \quad (10.8)$$

where λ is the wavelength of the radiation and A is the effective area of the aperture. We discuss modes below. This relation, which may be derived from the Fraunhofer integral and conservation of energy (as in Born and Wolf 1980; Sect. 8.3.3.), is a generalization of familiar results from diffraction theory. For example, Airy’s famous expression that the angular diameter of the

²We follow common notation but note that in  10.2, θ is a coordinate on the sky; in  10.4, θ is the angular measure of the beam profile with a $\theta = 0$ corresponding to the beam peak; and in  10.5, θ is the angular separation between two pixels on the sky.

spot size from a uniformly illuminated aperture is $2.44\lambda/D$, where D is the aperture diameter, is equivalent to (10.8).³ The relations above give a handy conversion between the system's effective aperture, the angular extent of the beam and the frequency of interest for single-mode optical systems. Combining (10.7) and (10.8) we obtain

$$\theta_{1/2,\text{rad}} = 1.06\lambda/D, \quad (10.9)$$

where D is the effective illumination on the primary reflector.

In the above treatment, we brought in the concept of a single propagating mode of radiation. A mode is a particular spatial pattern of the electromagnetic field. Radiation propagation in a rectangular waveguide of height $a/2$ and width a gives a familiar example. For frequencies less than a cutoff, $\nu_c < c/2a$, no electromagnetic radiation can propagate down a waveguide of length longer than a few λ . For $c/2a < \nu_c < \sqrt{5/4}c/a$, only the TE₁₀ mode of radiation propagates; at frequencies just above c/a , the TE₁₀ and TE₀₁ can propagate. Above $\sqrt{5/4}c/a$, the TE₁₀, TE₁₁, and TM₁₁ modes are free to propagate. With the geometry of a cylindrical waveguides of diameter d , the lowest frequency mode is the TE₁₁ (which supports two polarizations) with a cutoff frequency $\nu_c = c/1.7d$, and the next modes are TM₀₁ and TE₂₁ which turn on at frequencies that are 1.31 and 1.66 higher, respectively, than the lowest.

Experimentally, the selection for operating in a single mode is typically achieved by having a waveguide somewhere along the light path, typically at the entrance to the detecting element.⁴ The waveguide is essentially a high-pass filter, selecting the lowest frequency that can pass through the system. An additional low-pass filter then rejects frequencies at which the second and higher modes are propagating. Experimenters have been using single modes because these systems have particularly well-behaved and calculable beam patterns. If a second mode were added, say by operating at a higher frequency so that both the TE₁₁ and TM₀₁ propagated (for a cylindrical waveguide), one would receive more signal, an advantage, but the beam pattern of the combination of modes would be different, likely more complex compared to the single-mode illumination, and there would likely be increased spill over the edge of the primary.

Consider a radio receiver that observes a diffuse Planckian source of temperature T through a telescope. The surface brightness is given by

$$S_\nu(T) = \frac{2h\nu^3}{c^2(e^{h\nu/kT} - 1)} \rightarrow \frac{2\nu^2}{c^2}kT, \quad (10.10)$$

where h is Planck's constant, k is Boltzmann's constant, and S_ν is measured in W/m²srHz. The expression on the right is the surface brightness in the Rayleigh-Jeans limit. The power that makes it through to the detector is given by

$$P = \frac{1}{2} \int_{\Omega} \int_{\nu} \epsilon(\nu, \theta, \phi) A_e(\nu) S_\nu(\theta, \phi) B(\nu, \theta, \phi) d\Omega d\nu, \quad (10.11)$$

where the factor of 1/2 comes from coupling to a single polarization, A_e is the effective area, and ϵ is the transmission efficiency of the instrument. For clarity of discussion, we will henceforth

³The Airy beam profile is given by $B(\theta) = [2J_1(x)/x]^2$ where $x = \pi D \sin(\theta)/\lambda$ and J_1 is a Bessel function. The value of $1.22\lambda/D$ is the angular separation between the maximum and the first null. For small angles, $\theta_{1/2} = 1.03\lambda/D$. The total solid angle is $2\pi \int B_n(\theta) \sin(\theta) d\theta$. To make the integral simple and avoid considering the difference between projecting onto a plane versus a sphere, we consider the limit of small θ . Then, $\Omega = 8\lambda^2/\pi D^2 \int_0^\infty [J_1(\pi D x/\lambda)]^2 x^{-1} dx = \lambda^2/A$.

⁴In a close packed array, this may be approximated having the pixel size smaller than λ . Such a spatial mode would support two polarizations.

assume that the transmission efficiency is unity. If S_ν is uniform across the sky, we are in the Rayleigh-Jeans limit ($h\nu \ll kT$) and A_e and B_n are relatively independent of frequency over a small bandwidth (commonly achieved), then

$$P = \frac{1}{2} \int_\nu A_e(\nu) 2 \frac{\nu^2}{c^2} kT \int_\Omega B(\nu, \theta, \phi) d\Omega d\nu = kT \int_\nu \frac{A_e(\nu) \Omega}{\lambda^2} d\nu = kT \int_\nu d\nu = kT \Delta\nu. \quad (10.12)$$

Thus, each mode of radiation delivers $kT\Delta\nu$ of power to the detector. If there is a second mode in the system that is supported in this bandwidth, then it also contributes $kT\Delta\nu$ of power. It is possible, even likely, that different modes are supported over different but overlapping bandwidths.

Increasing the amount of celestial power on one's detector is an advantage when trying to detect a faint signal like the CMB. The trade-off is between control of the optical properties of the system and collecting power onto the detector. Note that using a larger telescope does not increase the detected power if one detects only a single mode. A larger telescope merely increases the resolution. In a bolometric system, one can to a certain extent control the number of modes that land on the detector. For example, one can place the absorbing area at the base of a "light collector" or Winston cone (Welton and Winston 1978). An approximation to the number of modes in the system is then found by beam mapping to determine Ω_B , measuring the pass-band to find the average wavelength λ_a , and dividing by the collecting area of the input optics. This gives the number of modes as $\alpha_m = A\Omega/\lambda_a^2$. This is only an approximation because it assumes knowledge of the aperture distribution (for the collecting area) and that all modes couple to the detector with the same efficiency. We use α because often this quantity is not an integer. Although formally modes come in integer sets, not all modes couple equally to the detector output. In the early days of CMB bolometry, multimoded systems were often used. As detectors became more sensitive, the field moved toward single-moded bolometric systems as pioneered in the White Dish experiment (Tucker et al. 1993). This led to more precise knowledge of the beams. To a good approximation, the current generation of bolometric CMB instruments all operate single-moded (with the first mode of propagation). However, there are modern examples of multimoded systems though they are not used for the primary CMB bands. They include the 345-GHz band on BOOMERANG (Jones 2005) and Planck's 545- and 857-GHz bands (Ade et al. 2010; Maffei et al. 2010) where there are just a few modes. In these cases, the coupling of the radiation in the bolometer's integrating cavity is in practice not possible to compute accurately. Interest in multimoded systems has returned with at least one satellite proposal for an instrument called PIXIE for measuring the CMB polarization in a massively overmoded system (Kogut et al. 2011). The PIXIE concept is based on the observation that the signal improves as the number of modes, n_m , but the noise degrades only as $\sqrt{n_m}$ in the photon-limited noise regime (see below). Thus, S/N improves as $\sqrt{n_m}$.

1.4 Noise

The choice of an optical system and its location is intimately connected with the desired noise performance. There are a number of contributing noise sources that depend on the type of detector, how it is biased, and on its environment (see, e.g., Mather 1982; Pospieszalski 1992). For this chapter, we concern ourselves primarily with the photon noise from the sky because it sets the ultimate detection limit. We first consider bolometric or "direct" detectors which

detect the total power and destroy all phase information in the incident field. Equation 10.13 gives the photon noise power on the detectors per mode (e.g., Zmuidzinas 2003) as

$$N^2(\nu)\tau = \frac{\Delta\nu}{\eta(\nu)(k\Delta\nu)^2} (h\nu)^2 n(\nu)[1 + \eta(\nu)n(\nu)], \quad (10.13)$$

where τ is the integration time, $\Delta\nu$ is the bandwidth, $n(\nu)$ is the occupation number (power in a mode divided by $h\nu$), and $\eta(\nu)$ is the quantum efficiency which we take to be unity. We have approximated integrals by multiplying by a bandwidth of $\Delta\nu$. Because each mode of radiation delivers a power of $kT\Delta\nu$, one may convert from the $W s^{1/2}$ to $K s^{1/2}$ by dividing by $k\Delta\nu$. The left-hand term in the expression is the Poisson term, and the right-hand term accounts for the correlations between the arrival times of the photons. When there are multiple modes in the system, one cannot simply assume that the above holds for each mode. One must take into account the correlations between the photon noise in each mode (Lamarre 1986; Richards 1994; Zmuidzinas 2003).

For coherent detector systems, one first amplifies the incident electric field while retaining phase information. After multiple stages of amplification, mixing, etc. one at last records the power in the signal. Because the amplitude and phase are measured simultaneously, and these quantities do not commute, quantum mechanics sets a fundamental noise limit of $N\sqrt{\tau} = h\nu/k\sqrt{\Delta\nu}$. In practice, the best systems achieve three times the quantum limit over a limited bandwidth and in ideal conditions. A good estimate of the noise limit is

$$N(\nu)\sqrt{\tau} = \frac{3(h\nu/k) + T_{\text{sky}}}{\sqrt{\Delta\nu}}, \quad (10.14)$$

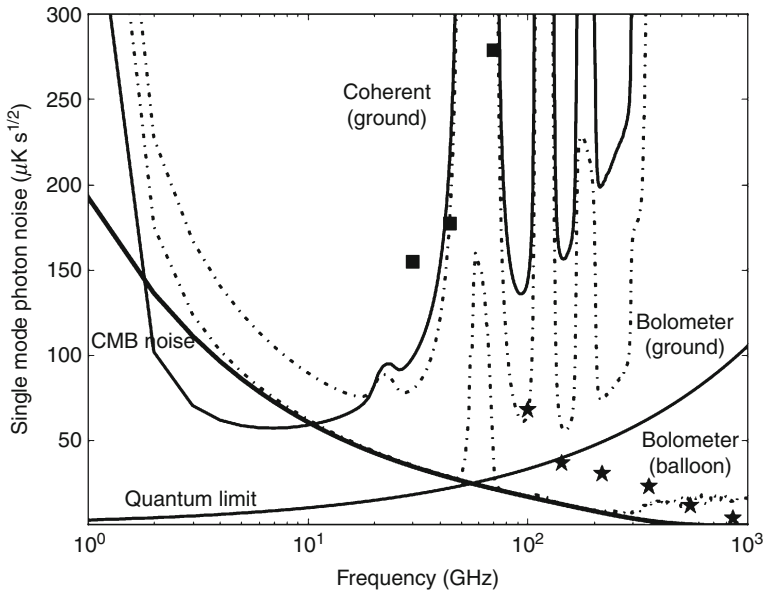
where T_{sky} is the antenna temperature of the incident radiation.

In Fig. 10-3, we show the noise limit for a single-moded detector with 20% bandwidth at a high-altitude ground-based site (e.g., the South Pole or the Atacama Desert) and at a typical balloon altitude of 36 km. We also show the noise level for the current generation of detectors on the Planck satellite. The quoted sensitivities for the bolometric detectors (Planck HFI Core Team et al. 2011) (two polarizations combined and all noise terms in the high-frequency limit) are adjusted up to a 20% reference bandwidth. The quoted sensitivities for the coherent detectors (Mennella et al. 2011) (all noise terms in the high-frequency limit) have been adjusted down to a 20% reference bandwidth.

Advances in bolometric detectors have reached the point where the intrinsic noise is near the noise limit set by the photon noise at a high-altitude site. Thus, to improve sensitivity, one wins more quickly by adding detectors as opposed to improving the detector noise. This is one of the motivations behind large arrays of detectors, and their associated large fields of view. One can also win by increasing the number of modes.

1.5 Polarization Terminology

There is a well-developed terminology for describing polarization. Imagine a telescope beam that points at a single position on the sky and feeds a detector that can measure the amplitude and phase of a partially coherent electric field. Because the electric field is a vector in a plane, it can be completely specified by measuring its horizontal, $E_x(t)$, and vertical, $E_y(t)$, components at each instant. To measure the intensity of the field, one averages the detector outputs over time.



■ Fig. 10-3

Photon noise from 1 to 1,000 GHz for a single mode of radiation for a 20% bandwidth in frequency. The CMB noise (thick solid) is for a region of sky without any other foreground emission. It sets a fundamental limit over most of this frequency range (the far infrared background, not shown, sets the limit near 1 THz). Noise from atmosphere (Chile or South Pole, zenith angle of 45°) for bolometers (dash dot) is lower than for coherent receivers (solid) except below ~20 GHz. Bolometers on balloon (dash dot) are limited by CMB noise between 80 and ~200 GHz. The atmospheric noise shown is due to thermal emission and does not include contributions from turbulence, changes in column density, or water vapor, which can increase the noise many-fold. Also shown are reported coherent receiver (square) and bolometer (star) noise for the Planck satellite, both adjusted as discussed in the text and both for total intensity. The Planck bolometers are close to the fundamental noise limit

The polarization properties of the field, assuming they are relatively constant, are completely specified with the coherency matrix:

$$\begin{pmatrix} \langle E_x E_x^* \rangle & \langle E_x E_y^* \rangle \\ \langle E_y E_x^* \rangle & \langle E_y E_y^* \rangle \end{pmatrix} \propto \frac{1}{2} \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \frac{1}{2} \begin{pmatrix} Q & U \\ U & -Q \end{pmatrix} + \frac{i}{2} \begin{pmatrix} 0 & V \\ -V & 0 \end{pmatrix} \quad (10.15)$$

where the “*” denotes a complex conjugate and the average is taken over time. The coherency matrix can also be represented by means of Stokes parameters I , Q , U , and V , as shown in the right-hand side of (10.15). The polarization has the symmetries of a spin-two field. The proportional sign indicates that the Stokes parameters, which represent intensities, are reported in Kelvins. The total intensity in the radiation is the trace of the matrix. Stokes Q is the intensity of the horizontal minus the vertical. Stokes U is the in-phase correlation between the two components of the field minus the 180° out-of-phase correlation. If the incident radiation was pure Stokes Q and one rotated the field by 45° in the x - y plane, then the output would be pure U . Stokes V measures circular polarization. However, the CMB is expected to be only linearly

polarized. (See Zaldarriaga and Seljak (1997) and Kamionkowski et al. (1997) for a discussion and formalism.) Although deviations from this prediction are of great interest, they are beyond the scope of this chapter.

As the beam is scanned across the sky, one makes maps of I , Q , and U . Of course, the values of Q and U depend on the specification of a coordinate system. However, the Q and U maps may be transformed into E-modes and B-modes. The advantage of these modes is that they are independent of the coordinate system and, for the CMB, are directly related to different physical processes in the early universe (Kamionkowski et al. 1997; Zaldarriaga and Seljak 1997). The E-modes correspond to a spin-two field with no curl and originate primarily from density perturbations in the early universe. This E-mode signal has been detected by a number of instruments. The B-modes correspond to a spin-two field with no divergence and can originate from tensor-type physical processes such as gravity waves that are predicted to have been generated by an inflationary epoch as close as 10^{-35} s after the big bang. To date the B-mode signal has not yet been discovered. If the B-modes were of sufficient amplitude to be detected, their impact on cosmology and physics would be enormous. Not only would the discovery significantly limit the number of models that could describe the early universe, but it would mark the first observational evidence of gravity operating on a quantum scale.

At large angular scales, $\ell \lesssim 100$, B-modes may result from inflationary gravitational waves and from galactic foreground emission. At higher ℓ multipoles, the primary contribution to the B-mode spectrum is from E-modes being gravitationally lensed so that they produce a B-mode component. The level of primordial (or inflationary) B-modes is quantified in terms of a parameter r , the ratio of the variance of density perturbations to tensor perturbations. Predictions for r vary over many orders of magnitude. Currently, observations give $r < 0.21$ (95%) (Keisler et al. 2011), a limit coming from *temperature* anisotropy and other cosmological probes (rather than polarization).⁵ When translated into temperature units in \blacktriangleright Fig. 10-2, this becomes a faint $\mathcal{B} \lesssim 150 \times 10^{-9}$ K for $\ell \sim 90$. The experimental challenge is to make accurate and precise polarization measurements at the level of few tens of nano-K.

2 Ground- and Balloon-Based Systems

In this section, we guide the reader through the set of considerations facing a designer of a CMB telescope. Once a resolution is chosen, one must consider whether to use a reflective or refractive system, a combination, or perhaps only a feedhorn. The information we provide is informed by the history of the field. We focus on a number of core design elements some of which have found use in multiple CMB experiments. Perhaps the largest difference in telescope design between CMB and other applications is that in CMB work the edge tapers on ambient temperature optics are kept low. We discuss the advantages and disadvantages of working on a balloon-borne platform. After the turn of the millennium, emphasis in the field turned toward large-throughput systems and polarization-sensitive experiments; both topics are discussed toward the end of the section.

⁵The inflation-generated gravity waves also contribute to the temperature anisotropy and thus can be constrained by such measurements.

2.1 General Considerations

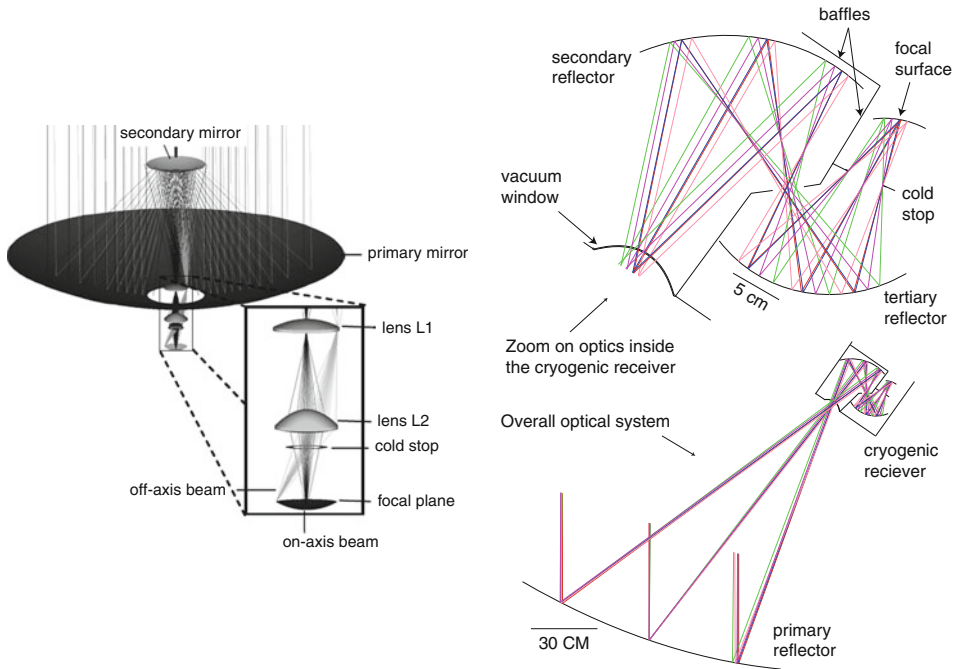
Following from the sky toward the detector, CMB optical hardware typically includes some or all of the following elements: reflectors, lenses, band-defining filters, amplifiers, detectors, and feedhorns or antennas. The order of some of these elements may vary somewhat. Intermediate elements, such as a vacuum window, are generally not considered part of the optical system although they do have to be considered as an optical element during the design of the system.

As discussed in [Sect. 1](#), the throughput is an overall measure of the amount of light that can be collected by the optical system. However, for a given aperture size and therefore for telescopes with the same resolution, the performance of the optical system is more usefully measured in terms of the “diffraction-limited field of view” (DLFOV), which is that portion of the focal surface across which the optical performance is diffraction limited. Two related measures are commonly used to determine whether the optical performance is diffraction limited: the Strehl ratio and the *rms* wavefront error. A system that provides a Strehl ratio larger than 0.8 (*rms* wavefront error that is less than $\lambda/14$) at a particular field point is generally considered diffraction limited at that field point. The minimum DLFOV is the area on the focal plane with a Strehl ratio larger than 0.8. However, some optical systems are optimized for higher Strehl ratios and therefore define the DLFOV as a smaller area enclosing a higher value.

The large majority of balloon- and ground-based CMB telescopes until the early 2000s illuminated the sky with either feedhorns or employed a combination of feedhorns and reflective optical systems. At the end of this chapter, we present a table of instruments that have published measurements of the primary CMB polarization or anisotropy. In the simplest feedhorn-only system, the far-field beam shape, its angular size, and the bandwidth are determined by the shape of the feedhorn and the waveguide components attached to it. The theory and design of feedhorns of various types is quite mature, is discussed in a number of publications, and is subject of ongoing research (Clarricoats and Olver 1984; Olver et al. 1994; Balanis 2005). We can use an approximate empirical relation for corrugated feeds, $\theta_{1/2,\text{deg}} = 90\lambda/D$, to show that practical considerations limit optical systems that use only feeds to have coarse resolution. A horn with a reasonably large effective aperture diameter of ~ 15 cm produces a beam size of $\sim 6^\circ$ and $\sim 1^\circ$ at 30 and 150 GHz, respectively. More complete studies show that it is difficult to produce a beam with $\theta_{1/2} < 2.5^\circ$ at 150 GHz with a corrugated structure of reasonable volume and tolerances (Lin, 2009, private communication for a study of a “feed farm” satellite concept available from http://phy-page-g5.princeton.edu/~page/cmbpol_feedfarm.pdf).

There are several design choices for reflector-based systems. In a “centered” (or sometimes “on-axis”) optical system, the reflectors are made from central portions of the typically conic sections of revolution that describe the surface shape. As a consequence, such systems naturally have central obscurations. Examples include the White Dish telescope (Tucker et al. 1993) and the QUAD optical system (Hinderks et al. 2009), shown in [Fig. 10-4](#), which consisted of a 1.2-m and 2.6-m parabolic primaries, respectively, and hyperbolic secondary in a Cassegrain configuration (QUAD also used lenses internal to the cryostat as part of the optical train).

In a decentered (sometimes also called “off-axis,” or “offset”) optical system off-axis portions of the conic sections are used, the reflectors are not centered on each other’s axis of symmetry, and there is no self-obscurations by the reflectors. An example of a decentered optical system of



■ Fig. 10-4

Examples of centered (*left*) and decentered (*right*) CMB optical systems. The centered, two-reflector Cassegrain system is that of the ground-based QUAD instrument (Figure from O'Sullivan et al. (2008)). The primary reflector diameter was 2.6 m. The optical elements after the secondary, shown in the inset, are all maintained at a temperature of 4 K. The decentered, three-reflector Gregorian system is that of the MAXIMA (Rabii et al. 2006) balloon-borne experiment. The primary is an off-axis section of a parabola with a diameter of 1.3 m. Two additional reflectors were maintained inside the cryogenic receiver (only part of which is shown) at a temperature of 4 K; see zoom

the MAXIMA experiment (Hanany et al. 2000) is shown in Fig. 10-4. For fixed entrance aperture diameter, centered systems are more compact compared to decentered systems. However, they have lower aperture efficiency and are more prone to scattering of radiation to side lobes caused by either diffraction from the edges of the central obscuration or by beam-intercepting supports of the secondary reflector(s). Decentered systems that have an intermediate focal point are typically easier to baffle compared to centered systems and are thus less prone to stray radiation. Most CMB telescopes to date use decentered systems; see Table 10-1.

Practical considerations limit the use of a refractive-only system. One is the weight of the lens. The density of ultrahigh molecular weight polyethylene, a popular material for millimeter wave lenses because of its low absorption, is 0.95 g/cm^3 , making a 1-m-diameter lens weigh around 100 kg. In contrast, the 1.3-m-diameter MAXIMA reflector weighed 11 kg, and the 1.4-m-diameter WMAP primary weighed only 5 kg. The Archeops primary, which was $1.8 \text{ m} \times 1.5 \text{ m}$, and was made of 6061 aluminum (rather than specialty materials, as the former two examples), weighed less than 50 kg. Another advantage of reflectors is their achromaticity. Many CMB instruments operate with multiple frequencies simultaneously, which facilitates the

■ Table 10-1
 CMB polarization and anisotropy experiments with comments on their optical and detector configurations. Much of the information about experiments prior to 2000 is adapted from "Finding the Big Bang" (Peebles et al. 2009). We include only instruments with astrophysical results as this indicates some level of the maturity of the design. Except for Planck, the citations are for the first astrophysical result from the instrument

Experiment	Type ^a	N _{feeds} ^b	N _{det} ^c	Optical design	Plat. ^d	Reference
Isotropometer	Dicke switched	2	1	Feed	Gnd	Wilkinson and Partridge (1967)
Stanford	Dicke-switched	2	2	Parabola	Gnd	Conklin and Bracewell (1967)
Crawford Hill	Maser	1	1	Horn/reflector	Gnd	Wilson and Penzias (1967)
Aerospace	Coherent	2	2	4.6 m telescope	Gnd	Epstein (1967)
White Mountain	Coherent	2	2	Dicke-switch two feeds	Gnd	Conklin (1969b)
Ratan	Coherent	1	1	Parabolic	Gnd	Pariiskii and Pyatunina (1971)
KaDip	Dicke-switched	2	2	Feeds	Gnd	Boughn et al. (1971)
XBal	Dicke-switched	2	2	Feeds	Bal	Henry (1971)
NRAO-P	Parametric amp	2	2	140 ft Greenbank Centered Cass.	Gnd	Parijskij (1973)
Goldstone	Maser	1	1	64 m Goldstone Centered Cass.	Gnd	Carpenter et al. (1973)
Parkes	Correlation	2	2	64 m Parkes Centered Cass.	Gnd	Stankevich (1974)
U2	Dicke-switched	2	2	Two corrugated feeds	Plane	Smoot et al. (1977)
Testa-Griga	Bolometers	1	1	Cass.	Gnd	Caderni et al. (1977)
Greenbank-R	Parametric amp	2	2	Cass.	Gnd	Rudnick (1978)
MIT	Bolometers	2	2	Two 0.4 m sph refl + flat + lightpipe	Bal	Muehlner (1977)
KKaQBal	Dicke-switched	2	2	Feeds	Bal	Cheng et al. (1979)
PolCMB	Coherent (P)	1	1	First pol	Gnd	Nanos (1979)
KPRO	Coherent	1	2	11 m Cass.	Gnd	Partridge (1980)
Convair	Bolometer	1	1	Lens/feed on FTS with chopper	Plane	Fabbri et al. (1980b)
DBal	Bolometer	1	1	Lens/feed	Bal	Fabbri et al. (1980a)
NRAO91	Coherent	2	2	91 m NRAO	Gnd	Ledden et al. (1980)
OVRO40	Coherent	2	2	OVRO 40 m	Gnd	Seielstad et al. (1981)
GBank-UW	Coherent	1	1	140 ft GB	Gnd	Uson and Wilkinson (1982)

Table 10-1
(Continued)

Experiment	Type ^a	N _{feeds} ^b	N _{det} ^c	Optical design	Plat. ^d	Reference
MaserBal	Maser	2	1	Dicke switch	Bal	Fixsen et al. (1983)
WBal	Mixer	1	1	Dicke-switched chopper	Bal	Lubin et al. (1983)
JodrellBank	Coherent	2	2	Prime focus of 100 ft MkII telescope	Gnd	Lasenby and Davies (1983)
Relikt	Parametric amp	2	2	Two feeds	Sat	Strukov and Skulachev (1984)
NCP	Coherent	2	2	Dicke switch w/ feed-fed parabolas	Gnd	Mandolesi et al. (1986)
Tenerife	Coherent	2	2	Two feeds with chopping plate	Gnd	Davies et al. (1987)
IAB-I	Bolometer	1	1	1 m parabola	Ant	dall'Oglio and de Bernardis (1988)
MITBal2	Bolometer	2	4	Horns and chopper	Bal	Halpern et al. (1988)
OVRO	Maser	2	1	OVRO 40 m	Gnd	Readhead et al. (1989)
SKInt	Mixer	2	2	Two feeds	Int	Timbie and Wilkinson (1990)
FIRS	Bolometer	1	4	Single cryogenic horn/lens	Bal	Page et al. (1990)
ARGO	Bolometer	1	4	1.2 m centered Cassegrain	Bal	de Bernardis et al. (1990)
SP/ACME	SIS Mixer	1	1	1 m decentered Gregorian	Ant	Meinhold and Lubin (1991)
COBE	Coherent	10	12	Feeds	Sat	Smoot et al. (1991)
SP/ACME	HEMT	1	1	1 m decentered Gregorian	Ant	Gaier et al. (1992)
19 GHz	Maser	1	1	Feed plus lens	Bal	Boughn et al. (1992)
MAX	Bolometer	3	3	Same as SP/ACME	Bal	Alsop et al. (1992)
IAB-II	Bolometer	1	1	0.45 m decentered Cassegrain	Ant	Piccirillo and Calisse (1993)
White Dish	Bolometer	1	1	1.2 m centered Cassegrain	Ant	Tucker et al. (1993)
SASK	HEMT	1	1	1.2 m off-axis parabola	Gnd	Wollack et al. (1993)
MSAM	Bolometer's	1	4	1.4 m decentered Cassegrain	Bal	Cheng et al. (1994)
PYTHON	Bolometer's	4	4	off-axis parabola	Ant	Dragovan et al. (1994)
CAT	Coherent	3	6	Int	Gnd	O'Sullivan et al. (1995)
BAM	Bolometer	2	2	FTS with off-axis parabola	Bal	Tucker et al. (1997)
SuzIE	Bolometer	6	6	CSO	Gnd	Ganga et al. (1997)
IAC-BAR	Bolometer	4	4	0.45 m decentered para/hyper	Gnd	Piccirillo et al. (1997)
QMAP	HEMT	3	6	Feed + parabolic reflector	Bal	de Oliveira-Costa et al. (1998)

Toco	HEMT/SIS	5	8	Feed + parabolic reflector	Gnd	Miller et al. (1999)
JB-IAC	HEMTs	2	2	Feed plus reflector	Int	Dicker et al. (1999)
HACME	HEMT	1	1	Decentered Greg.	Bal	Staren et al. (2000)
Viper	HEMT	2	2	Decentered aplanatic Greg. with chopper	Ant	Peterson et al. (2000)
RING5M	HEMT	2	1	5.5 m/40 m OVRO Centered Cass.	Gnd	Leitch et al. (2000)
BOOMERANG	Bolometers	16	16	Decentered Greg. w/tert.+ lenses	Bal	de Bernardis et al. (2000)
MAXIMA	Bolometers	16	16	Decentered Greg.	Bal	Hanany et al. (2000)
PIQUE	HEMT (P)	1	2	1.2 m off-axis parabola	Gnd	Hedman et al. (2001)
POLAR	HEMT (P)	1	2	Cryo feed	Gnd	Keating et al. (2001)
CBI	HEMTs	13	13	Centered Cass.	Int	Padin et al. (2001)
DASI	HEMTs	13	13	Feeds	Int	Halverson et al. (2002)
Archeops	Bolometers	21	21	Decentered Greg.	Bal	Benoît et al. (2003)
COMPASS	HEMT (P)	1	2	2.6 m centered Cass.	Gnd	Farese et al. (2003)
VSA	HEMTs	14	14	Feeds + off-axis parabolas	Int	Grainge et al. (2003)
WMAP	HEMTs	20	40	Decentered Greg.	Sat	Bennett et al. (2003)
Acbar	Bolometers	16	16	Decentered aplan. Greg. w/ chopper	Ant	Kuo et al. (2004)
BEAST	HEMT	8	8	2 m decentered Greg.	Bal	Meinhold et al. (2005)
CAPMAP	HEMT (P)	16	32	7 m decentered Cassegrain	Gnd	Barkats et al. (2005)
MINT	Mixers	4	4	30 cm Cass.	Int	Fowler et al. (2005)
MAXIPOL	Bolometer (P)	16	16	Decentered Greg.	Bal	Johnson et al. (2007)
QUAD	Bolometer (P)	31	62	Centered Cass. w/lenses	Ant	Ade et al. (2008)
WMPol	HEMT (P)	4	3	2.2 m decentered Greg.	Ant	Levy et al. (2008)
SPT	TES bolometer	966	966	Decentered Greg. w/lens	Gnd	Staniszewski et al. (2009)
BICEP	Bolometer (P)	49	98	Centered refractive	Ant	Chiang et al. (2010)
ACT	TES bolometer	Planar	3,072	Decentered Greg. w/lenses	Gnd	Fowler et al. (2010)
QUIET	HEMTs (P)	110	110	Crossed Dragone	Gnd	QUIET Collaboration et al. (2011)
Planck	HEMT/bol (P)	47	74	1.6 m aplanatic Gregorian	Sat	Tauber et al. (2010a)

^a Detector technology, (P) indicated a design specifically for polarization

^b Number of feeds

^c Number of independent detectors

^d Platform. *Gnd* ground, *Bal* balloon, *Sat* satellite, *plane* airplane, *Ant* Antarctica, *Int* interferometer

discrimination of foregrounds sources from the CMB signal. Eliminating the $\sim 4\text{--}30\%$ reflection per refracting surface, for polyethylene and silicon, respectively, requires antireflection coatings that operate over a correspondingly broad range of frequencies, assuming that the different frequencies share the same light train. In contrast, metal-based reflectors can have nearly unity reflectance between few MHz and several THz. To date, there have been less than a handful of CMB refractive-only systems. They used polyethylene lenses of ~ 30 cm diameter and hence operated at degree-scale resolution, and with one or at most two frequencies per optical train. An advantage of a refractive-only system is that it typically provides a large DLFOV with no obscurations and in a relatively compact package.


A common theme for all CMB telescopes is controlling and understanding the sidelobes. Invariably antenna patterns have nonvanishing response beyond the design radius of the optical elements. This “spillover” is quantified by the edge taper, which is the level of illumination at the edge of the optic relative to the center of the beam, typically quoted in dB. The best practice is to ensure that such spillover radiation finds itself absorbed on a cold surface with stable temperature. How cold and how stable? Sufficiently cold such that the total power coupled through the spillover is small compared to the total load on the detectors and stable compared to the timescales of the largest sky scans. Through scattering, some of the spillover radiation can find its way to the sky, potentially coupling astronomical sources that are away from the main beam, or to the ground causing spurious signals as the telescope scans the sky. Three techniques have been used most widely to control sidelobes, sometimes in combination with each other: (1) use of specifically shaped feedhorns to launch the beams from the detector element into the rest of the optics (or to the sky). As mentioned earlier, the theory of the beam patterns produced by feedhorns is quite developed, and measurements are in good agreement with predictions. Thus, one can design the antenna pattern to have only low levels of spillover. (2) Use of an aperture stop to control and define the illumination on the primary reflector. A millimeter-wave black and often cold aperture stop is placed at a location along the optical path that has an image of the primary. Adjusting the diameter of the stop effectively controls the illumination and edge taper on the primary optic. In addition, radiation on the wings of the beam, at levels below the edge taper on the stop, is intercepted by cold surfaces. (3) Use of shields and baffles to absorb and redirect spillover radiation.


2.2 Balloon- Versus Ground-Based Systems

High instantaneous sensitivity is the primary driver to mount a CMB experiment on a balloon-borne platform. This improved sensitivity over a ground-based system is a consequence of two distinct elements: lower atmospheric emission and higher atmospheric stability. There is significantly less atmospheric power loading at balloon altitudes compared to ground, leading to lower photon noise; see [► Figs. 10-1](#) and [► 10-3](#). At 150 GHz, the atmospheric loading is about 100 times smaller at balloon altitudes than on the ground. This implies that if balloon-based systems control detector noise and emission from the telescope, their fundamental noise limit could be the CMB itself. As a generic example, assume a telescope similar to MAX-IMA's (Rabii et al. 2006) with one ambient and two aluminum reflectors cooled to 4 K. For simplicity, let's assume a flat passband between 120 and 180 GHz. Since the instrument operates in the single-mode limit at an effective wavelength of $\lambda = 0.2$ cm, the throughput per feedhorn is $0.04 \text{ cm}^2 \text{ sr}$ ([► 10.8](#)). Thus, the expected power from the CMB incident on the telescope

is 0.9 pWatt.⁶ In comparison, the power emitted by a 250-K ambient (at float) temperature reflector is 0.6 pWatt. This value assumes the 150-GHz bulk emissivity of aluminum of 0.13%. The *two* cold reflectors contribute together a negligible 0.01 pWatt.

This example illustrates several issues. Emission from even a single ambient temperature reflector on a balloon platform is not negligible compared to the power from the CMB. Emission from a telescope with two ambient temperature reflectors, as in, for example, EBEX (Reichborn-Kjennerud et al. 2010), would give rise to loading higher than the CMB. These calculations assume the lowest aluminum emissivity, that is, the bulk emissivity. If the actual surface has an effective emissivity of 0.5%, because of, for example, a layer of contamination, the power produced by a single warm reflector is 2.1 pWatt, which is more than twice the power from the CMB.

An effective way to mitigate emission is to maintain optical elements at low temperatures. This technique is particularly useful for optical systems with small apertures. It has been used with the ground-based BICEP (Aikin et al. 2010) experiment and the balloon-borne FIRS (Page et al. 1994) and Arcade (Singal et al. 2011) instruments. Other experiments had one or two ambient temperature reflectors but maintained subsequent optical components at cryogenic temperatures; see, for example,  Fig. 10-4.

In addition to higher photon loading, *variations* in atmospheric emission on the ground, essentially due to emission from transiting clouds of water vapor, are a source of increased noise at low temporal frequencies. The $1/f$ knee of this extra noise, that is, the low-frequency point at which the higher frequency white noise level doubles, varies with observing site on Earth and with specific atmospheric conditions. The combination of low-frequency noise and higher loading makes CMB observations above ~ 250 GHz difficult from anywhere on Earth. Observations *at large angular scales* are a challenge at almost any frequency because of the spatial structure of the atmosphere.⁷ The only ground-based experiment to report anisotropy measurements at angular scales larger than $\sim 8^\circ$ ($\ell \lesssim 20$) is Tenerife (Davies et al. 1987; Watson et al. 1992), which operated at 10 and 15 GHz. As shown in  Fig. 10-1, atmospheric emission drops at lower microwave frequencies in part making the Tenerife measurement possible. The significantly lower atmospheric emission on a balloon-borne platform and in particular the absence of water clouds essentially eliminate atmospheric emission as a significant source of low-frequency noise.

The situation is different for measurements of the polarization of the CMB. One may think of a polarization measurement of Stokes Q or U as the difference in intensities of two polarization states.⁸ If the two measurements are done simultaneously, or quickly relative to the timescale of atmospheric turbulence, the measurement is immune to fluctuations in atmospheric emission *if* the atmosphere is not polarized. Zeeman splitting of oxygen in Earth's magnetic field (Weiss 1980) polarizes atmosphere emission near the strong oxygen lines at 60 and 118 GHz. However, it has been shown that the atmospheric linear polarization is negligible compared to

⁶1 pWatt = 10^{-12} W.

⁷Spatial turbulence in the atmosphere is parametrized in terms of a Kolmogorov spectrum (Tatarskii 1961; Church 1995; Lay and Halverson 2000) that depends on the spatial wave number q as either $q^{-11/3}$ or $q^{-8/3}$ depending on whether the turbulent layer is three- or two-dimensional. Thus at large angular scale, low q , atmospheric fluctuations can be quite large.

⁸For bolometric systems one can simply imagine the difference of two intensity measurements. For coherent or interferometric systems, the square law detector outputs the product of two differently polarized electric fields.

the levels expected for either CMB E- or B-modes (Keating et al. 1998). Circular polarization near the lines is significantly higher than linear (Hanany and Rosenkranz 2003), and conversion of atmospheric circular to linear polarization in the instrument is a source of concern. Calculations of the effect of atmosphere polarization as a function of *spatial scale* depend on knowledge of Earth's magnetic field over the corresponding spatial scales. Available information (Finlay et al. 2010) suggests that spatial variations in the field are limited to large angular scales, roughly above 10° . Thus, experiments probing polarization anisotropy at smaller angular scales should not be affected by atmospheric polarization. For this reason, there is a relative abundance of experiments probing polarization from ground-based instruments at frequencies below 250 GHz. Despite these comments, we note that large angular scale polarization ($\ell \lesssim 30$, $\theta \gtrsim 6^\circ$) has so far been measured only by the WMAP satellite.

While the higher instantaneous sensitivity on a balloon-borne platform is appealing, it comes with particular optical-system and total integration-time trade-offs. To date, the largest CMB reflector mounted on a balloon platform was the 2.2-m primary of the BEAST experiment (Meinhold et al. 2005). It was made with carbon-fiber technology and weighed 8 kg. A similar reflector made using aluminum would likely weigh in excess of 50 kg. Using significantly larger telescopes is challenging, and thus resolution has a practical upper limit for a balloon-borne platform. Although BEAST operated at 40 GHz with $\theta_{1/2} = 23'$, one can imagine using a similar telescope to achieve $\theta_{1/2} = 6'$ at 150 GHz.⁹ To take full advantage of the balloon platform, though, it would be advantageous to cool the large reflector to well below ambient temperature to minimize radiation loading of the detectors. So far, this has not been achieved. Finally, the duration of a balloon flight is ~ 1 day (for launches in North America) to ~ 20 days (for two circumnavigations in a long duration flight in Antarctica),¹⁰ and a balloon experiment can typically be launched once every 1–2 years. In contrast, ground-based observatories have the potential for significantly longer continuous integration times.

2.3 Arrays of Detectors and Increases in DLFOV

Improvements in detector sensitivity throughout the 1990s placed new demands on optical systems. The sensitivity of individual detectors approached the photon noise limit (see [Fig. 10-3](#)), earlier for ground-based experiments for which the atmosphere is a strong source of emission, and then even for balloon-borne payloads. Improved experiment sensitivity could only be realized by using arrays of detectors in the focal plane and thus by increasing the DLFOV. Small-sized focal plane arrays were used in the late 1990s by MAXIMA (16 elements), BOOMERANG (16 elements, Crill et al. 2003), QMAP (6 elements, de Oliveira-Costa et al. 1998), and Toco (8 elements, Miller et al. 1999). The detector elements were bolometers that used neutron transmutation-doped germanium (MAXIMA, BOOMERANG), high electron mobility transistors, or HEMTs (QMAP, Toco), and superconductor-insulator-superconductor, or SIS, mixers (Toco). The push for large focal plane arrays accelerated in the early 2000, after the launch of

⁹For example, the BLAST balloon payload (Pascale et al. 2008), which had frequency bands between 600 and 1,200 GHz, had a centered Cassegrain system with a 2-m aperture primary providing a resolution of $30''$ at the highest frequency.

¹⁰Currently, record duration for a science payload in an Antarctic flight is close to 42 days (Seo et al. 2008). NASA is developing capabilities for flights of 100 days.

WMAP, and when Planck was slated for launch later in the decade. The main focus in CMB studies evolved toward measurements of the faint polarization of the CMB and of small-scale anisotropy, and it was broadly recognized that significantly higher mapping speeds can only be achieved by implementing large arrays of detectors.

To achieve higher throughput, CMB telescope designers turned to on-axis refractive systems and to new reflective optical systems that included additional corrections of aberrations. In addition, modern detector arrays with hundreds to thousands of elements are primarily based on superconducting technologies (such as transition-edge sensor (TES) bolometers, e.g., Irwin and Hilton (2005)) and are thus fabricated on flat silicon wafers. This requires optical systems to have flat focal planes and, depending on the coupling to the detectors, can require that the focal planes are image-space telecentric, namely, that the focal surface is perpendicular to all incident chief rays. The use of large focal plane arrays also prompted designers to be more conservative in their definition of DLFOV and strive for Strehl ratio values higher than the accepted minimum of 0.8. Higher Strehl ratios are generically associated with smaller beam asymmetries, which simplifies data analysis by reducing variations in the window functions across the focal plane array. What Strehl value should one strive for? We are not familiar with a quantitative study of the effects of minimum Strehl ratio on specific beam asymmetry nor its effects on the systematic error budget for any experiment. Designers of large arrays have generally attempted to provide for Strehl ratios larger than 0.9.

The advent of large focal plane arrays has stimulated research and innovation into the specific coupling of the electromagnetic radiation into bolometric TES arrays. A variety of coupling approaches have been developed, including feedhorn, contacting or immersion lens, phased antenna, and filled focal plane arrays (e.g., Padin et al. 2008; O'Brien et al. 2008; Kuo et al. 2008; Niemack et al. 2008; Yoon et al. 2009). The optical coupling of the detector array can have a substantial impact on the total system throughput and can drive critical optical design decisions. Here we provide a brief overview of the trade-offs between filled and feedhorn-coupled focal planes following the analysis of Griffin et al. (2002). ➤ Section 3 provides an example comparison between the optical coupling techniques used in the ACT and SPT focal planes.

For an instrument with a fixed DLFOV that is fully populated with detectors, a filled focal plane array of bare detectors can provide ~ 3 times faster mapping speed of an extended source (like the CMB) and with $0.5F\lambda$ spacing¹¹ ~ 3.5 times faster mapping speed of point sources than a feedhorn-coupled array with $2F\lambda$ spacing. For an instrument that is limited by readout technologies to a fixed number of detectors, feedhorn-coupled arrays can provide the fastest mapping speed by increasing the FOV until the feedhorns are spaced by $\sim 2F\lambda$. In practice, large detector arrays to date have generally operated between these extremes with feedhorn-coupled arrays spaced between 1 and $2F\lambda$ and filled arrays spaced between 0.5 and $1F\lambda$.


The differences in instrument design requirements for filled focal plane arrays compared to feedhorn (or other beam-forming element) arrays are substantial. The Gaussian illumination of feedhorns provides well-understood beam properties (and potentially single-moded coupling) and strong rejection of stray light from sources outside the main beam, and the horn itself provides a Faraday enclosure around the detector. In contrast, filled detector arrays are exposed,

¹¹Note that parameterizing the detector spacing or the feedhorn aperture in units of the focal ratio, F , times the wavelength, λ , provides sufficient information to approximate the aperture efficiency, spillover efficiency, and other relevant optical quantities for estimating the mapping speed.


can couple to radiation from approximately π steradian, and therefore require a cryogenic stop cooled to ~ 1 K to prevent blackbody radiation within the cryostat from dominating the detector noise. Individual detectors in a filled focal plane are smaller and have lower optical loading than feedhorn-coupled detectors, making it more difficult to achieve photon-limited noise performance and requiring readout of more detectors. The great potential advantage of filled arrays is of course the opportunity to maximize the throughput and mapping speed of detector arrays that fill the available field of view.

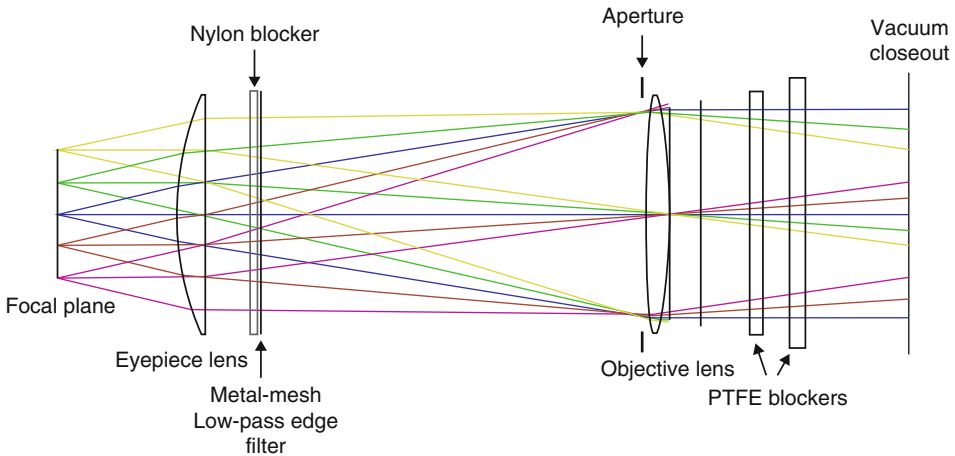
Another approach to maximizing the use of the DLFOV is to increase the optical bandwidth of the optics and have multiple detectors operating in different frequency bands within each focal plane element. This can be achieved by use of dichroic beam splitters in the optical path that illuminate independent detector arrays, or by optical coupling techniques being developed to make “multichroic” detectors by separating the frequency bands using superconducting filters integrated into the detector array (e.g., O’Brien et al. 2008; Schlaerth et al. 2010; McMahon et al. 2012). Advantages of multichroic detector arrays compared to dichroic beam splitters include more compact optics designs, fewer optical elements, a smaller detector array footprint, and the ability to fill the entire DLFOV in high optical throughput systems.

2.4 Refractor-Based Optical Systems

In refractive systems, the absence of a central obscuration, combined with the on-axis nature of the optics, provides large easy gains in DLFOV. For example, a NASA study for a CMB polarization space-based mission (Bock et al. 2008) presented an optical design that had an aperture of 30 cm, total throughput of $40 \text{ cm}^2\text{sr}$, Strehl ratio larger than 0.99 over a FOV of 15.3° in diameter, and resolution of 0.57° at 135 GHz. The optical performance was achieved with two polyethylene lenses that were pure conic sections. To our knowledge, BICEP was the first CMB experiment to *implement* a refractive-only system; the focal plane was indeed telecentric, as shown in  Fig. 10-5. The throughput was $34 \text{ cm}^2\text{sr}$ ¹² and it had resolution of 0.6° at 150 GHz (Yoon et al. 2006). The optical design approach of the BICEP system is also used in the ground-based Keck-Array (Sheehy et al. 2010) and the balloon-borne SPIDER (Filippini et al. 2010) instruments. They use five and six independent receivers, respectively, each of which is a fully refractive optical system similar to BICEP’s, to increase the total throughput of the entire system.

To achieve a combination of high throughput *and* high resolution, other experiments have implemented a combination of reflective fore-optics (e.g., primary and secondary reflectors) and refractive back-optics. The roles of the lenses is to further correct aberrations and to produce telecentricity. Cases in point are EBEX, Polarbear (Arnold et al. 2010), ACT, and SPT.

¹²Throughput calculations assume  10.6 taking the entrance aperture diameter and the total FOV available by the optical system of the experiment. A significantly smaller throughput value can be obtained by taking the throughput per detector element and multiplying by the number of detectors implemented. We opt for the first version because our primary interest in this chapter is in the overall optical design independent of the choice of detector spacing on the focal plane.



■ Fig. 10-5

The all-refractive optical system of BICEP (Figure is from Aikin et al. (2010)). The vacuum window is on the right. The two lenses and other filters are all maintained at a temperature of 4 K

2.5 Reflector-Based Optical Systems

The large majority of CMB telescopes to date are based on decentered reflecting systems.¹³ Most designers have found that systems with up to three powered reflectors¹⁴ are sufficient to provide the DLFOV necessary for their experiments. Minimizing the number of reflectors makes the system more compact, simpler to assemble, and easier to analyze in terms of misalignment errors.

With essentially any ray-tracing program, it is fairly straightforward to show that the DLFOV produced by a single reflector – typically, but not always an off-axis section of a parabola – is rather limited. The primary limiting aberration is astigmatism (Chang and Prata 2005). Nevertheless, a combination of feedhorn and a single reflector was used with the Saskatoon experiment and later for QMAP and Toco, which used the same telescope configuration and managed to pack six and eight element array of detectors in the focal plane, respectively. With two-reflector systems, the classical configurations are either a Cassegrain (parabolic primary and hyperbolic secondary) or a Gregorian (parabolic primary and elliptical secondary). We don't know of strong preferences between the two systems from an image quality point of view. However, the Gregorian system offers advantages in that the intermediate focus point between the primary and secondary reflectors is a convenient location for baffling (see, e.g., the decentered system in ◀ Fig. 10-4) and that for a fixed aperture size the Gregorian has been shown to be more compact (Brown and Prata 1994). Likely for these reasons most CMB experiments that use reflectors opt for the Gregorian configuration.


¹³Korsch (1991) and Love et al. (1978) give thorough reviews of reflecting optical systems.

¹⁴The term “powered” reflectors refers to reflectors with focusing properties, rather than flat reflectors used to only fold the path of the beam.

The primary source of aberrations in both Gregorian and Cassegrain systems are astigmatism and coma. Therefore, a good design starting point are those designs that provide some cancelation of either of these aberrations, at least at the center of the FOV. In an aplanatic Gregorian, making the primary slightly elliptical and the secondary a more eccentric ellipsoid than in the classical Gregorian cancels coma. The Planck optical system is based on an aplanatic Gregorian design.

Dragone has described designs that cancel astigmatism (D1) and both astigmatism and coma (D2) at the center of the FOV of either Gregorian or Cassegrain systems (Dragone 1982, 1983a). In both cases, Dragone uses the concept of a single equivalent paraboloid that has the same antenna pattern properties (at the center of the FOV) as the two-reflector system. He finds modifications to this equivalent system that cancel aberrations and then translates these modifications back to the two-reflector system. In a Gregorian D1 system, the modification is conceptually simple, it is a relative tilt between the axes of symmetry of the secondary ellipsoid and the primary parabola.¹⁵ In D2, Dragone derives additional corrections to the shape of the reflectors such as to cancel coma. Variants of D1 have become a popular starting point for several CMB designers. Hanany and Marrone (2002) compared the performance of several optical designs including the classical Gregorian, the aplanatic Gregorian, the D1, and the D2 and showed that the D2 provides the largest DLFOV.


To our knowledge, ACME was the first telescope based on a Gregorian Dragone design (Meinhold et al. 1993). The reflectors for EBEX and Polarbear are exact D1 designs. The WMAP design started from a D1 design and Planck's design is inspired by D1 in that it is an aplanatic but with the addition of the D1 tilt between the symmetry axes of the primary and secondary reflectors.


A third reflector, or equivalently refracting elements, are typically added to the base two-reflector system to achieve additional requirements of the optical system, such as to produce a distinct aperture stop, to further cancel aberrations and thus increase the DLFOV, or to make the focal surface telecentric. In MAXIMA, a cold aperture stop is placed at an image of the primary that is formed by the tertiary reflector; see  Fig. 10-4. In BOOMERANG, the tertiary reflector is the aperture stop. In ACT, EBEX, and Polarbear, lenses that reimage the primary form the cold stop. In all cases, the cold stop is used to control the illumination on the primary reflector.

A useful property of the D1 and D2 designs is that they cancel cross polarization at the center of the field of view. This property has been pointed out by earlier authors (see footnote 15). Minimum cross polarization is useful for telecommunication systems that can double the bandwidth by using two distinct polarizations with a single antenna. It is also useful for polarization-sensitive CMB experiments, but this usefulness is limited as will be discussed in the next section.

An innovative two-reflector system that provides a particularly large FOV is described by Dragone (1983b). This system, which has become known as the “crossed-Dragone”

¹⁵Graham (1973) first suggested introducing such a tilt in a decentered Cassegrain telescope to eliminate the cross polarization introduced by the asymmetrical configuration of the two reflectors. Subsequently, Mizuguchi and Yokoi (1974) and Mizuguchi and Yokoi (1975), and later others (e.g., Mizuguchi et al. 1976; Mizuguchi et al. 1978; Dragone 1978), made the idea more quantitative, expanded it to a Gregorian system, and to a system with more than two reflectors. (Some papers are published with Mizuguchi in place of Mizuguchi.) The primary motivation in these studies remained the elimination of cross polarization at the center of the FOV. Thus, a decentered Cassegrain or Gregorian optical system with a tilt between the axes of symmetry of the primary and secondary is sometimes referred to as a “Mizuguchi-Dragone telescope.” In a series of publications in the early 1980s, Dragone analyzes *aberrations* in decentered reflecting system. It so happens that the tilt that cancels cross polarization also cancels astigmatism.

configuration, began to be adopted by some CMB experimenters in the mid-2000.¹⁶ However, the first use of the general configuration for the CMB was by the IAB experiment (Piccirillo 1991; Piccirillo and Calisse 1993). The crossed-Dragone system has a parabolic primary and hyperbolic secondary and is thus essentially a Cassegrain system. It minimizes both astigmatism and coma near the center of the FOV without introducing spherical aberrations. An exact implementation of the design is shown in  Fig. 10-6. Variants of the crossed-Dragone system and detailed design procedures are described by Chang and Prata (1999). In addition to the large DLFOV provided by this system, the focal plane is nearly telecentric and the cross polarization is small. Tran et al. (2008) compared the performance of the Gregorian Dragone (of the D1 variety) and crossed-Dragone systems. They found that the DLFOV provided by the crossed-Dragone is about twice that provided by the Gregorian Dragone.

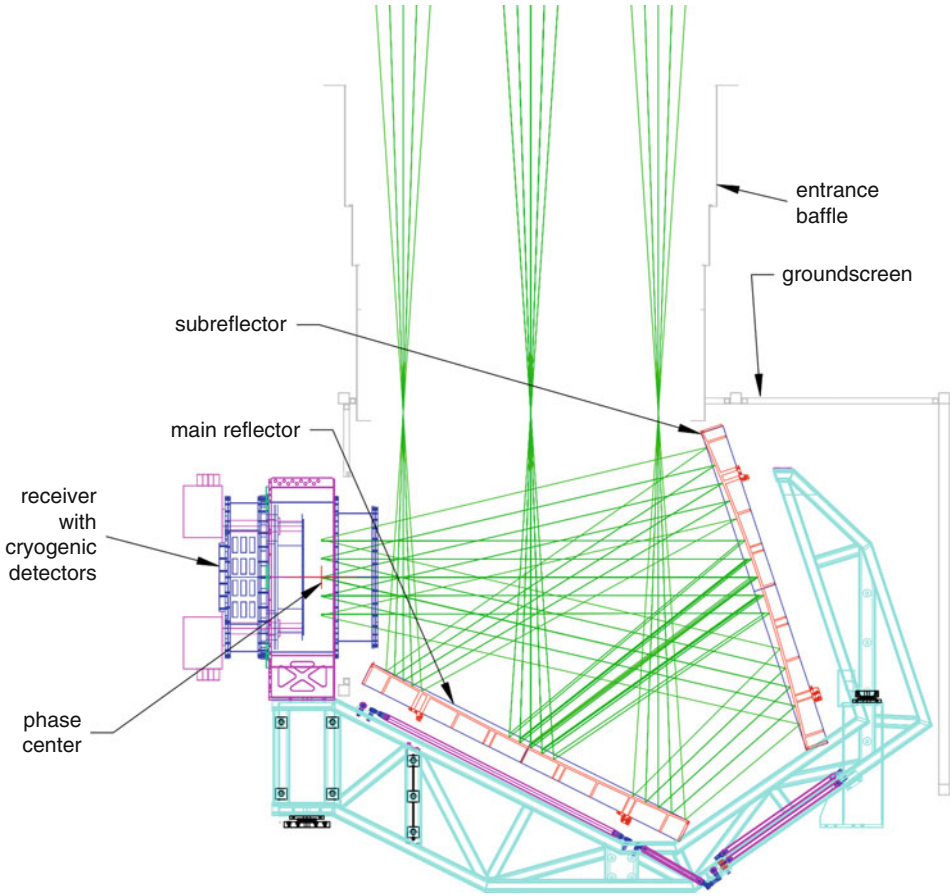
A crossed-Dragone system was developed for CLOVER (Piccirillo et al. 2008), a UK-based CMB experiment that has since been canceled. It was also proposed as an alternative optical system in an initial concept study, called EPIC, for a future CMB polarization satellite (Bock et al. 2008). In a subsequent round, the team had baselined the crossed-Dragone as its flagship design (Bock et al. 2009; Tran et al. 2010). A very large throughput of $908 \text{ cm}^2 \text{ sr}$ was obtained at 100 GHz with an aperture diameter of 1.4 m. The ambitious design accommodated a total of 11,000 detectors in an elliptical focal plane with long and short axes of 1.5 and 1 m, respectively, and Strehl ratios larger than 0.8 over this entire, nearly flat focal surface. The first CMB experiment that has made measurements with a crossed-Dragone system was QUIET (Imbriale et al. 2011); see  Fig. 10-6. The telescope throughput was $\sim 530 \text{ cm}^2 \text{ sr}$ with a 1.4-m primary reflector, and 0.22-deg resolution at 90 GHz. The ground-based ABS (Essinger-Hileman 2011), which operates at $\sim 145 \text{ GHz}$, uses smaller, 60-cm reflector primary; both primary and secondary are maintained at LHe temperature. In the ABS design, the entrance aperture is somewhat sky-side of the primary reflector, rather than on the reflector itself. This modification, which was later adopted by the EPIC team, is used to better define the illumination on the primary reflector and to control side lobes, with minimal penalty in optical performance.

Despite its appealing features, the crossed-Dragone system has several challenges which are described in Tran and Page (2009). One challenge is that the system's aperture stop is at the very front effectively limiting the resolution for systems in which this stop needs to be cold. Another challenge is the compactness of the system. Both the secondary reflector and focal surface are very close to the incoming bundle of rays causing diffraction side lobes on the edges of the reflectors and leading to physical packing difficulties when implementing a cryogenic focal plane. Tran et al. (2010) analyze the issue of side lobes in more detail in the context of the EPIC Mission Concept report.

2.6 Polarization Properties

Interest in measurements of the polarization of the CMB has motivated interest in the polarization properties of telescopes and other optical components. This is a relatively new area of research, for which only limited experience is available.

¹⁶In a parallel development, Vokurka (1980) proposed a 'crossed configuration' made of two cylindrically parabolic mirrors for an improved compact test range antenna. This crossed concept was expanded by Dudok and Fasold (1986) - also in the context of compact test range antenna - to a Cassegrain system similar to the one proposed earlier by Dragone, albeit without the various tilts and shape changes that reduce aberrations.



■ Fig. 10-6

The crossed-Dragone telescope of the QUIET experiment. The primary reflector is 1.4 m in diameter; in the ray limit it is the limiting aperture in the system. The system was designed using physical optics by propagating beams from the focal plane horns to the sky. An apparent entrance aperture is formed sky side of the primary, even though there is no physical aperture stop there. The entrance baffle intercepts side lobes that are inherent to the crossed-Dragone design, as described in the text

The susceptibility of an instrument to polarimetric systematic errors depends on the experimental approach to the polarization measurements, and a detailed discussion is beyond the scope of this chapter. However, considering the optical system alone, one can generalize as follows: systematic errors are minimized if the antenna pattern is completely independent of the polarization state being probed, if such an antenna pattern is independent of field position in the focal plane, and if the incident polarization orientation is not altered by the optical system. These desirables are never fully satisfied and a body of literature has evolved around quantifying the systematic errors induced by nonidealities (e.g., Hu et al. 2003; O’Dea et al. 2007; Shimon et al. 2008; Su et al. 2011).

As a starting point, one models the antenna patterns in each of the two polarization states $i = a, b$ as an elliptical Gaussian (Hu et al. 2003)

$$B_i(A_i, \hat{n}, \vec{b}_i, e_i) = A_i \exp \left[-\frac{1}{2\sigma_i^2} \left(\frac{(n_1 - b_{i1})^2}{(1 + e_i)^2} + \frac{(n_2 - b_{i2})^2}{(1 - e_i)^2} \right) \right], \quad (10.16)$$

where \vec{b} is an offset between the beam center and a nominal direction \hat{n} on the sky, σ is the mean beam width, and e is the ellipticity. It has become common to quantify beam-induced errors in terms of differential gain g

$$g \propto \frac{A_a - A_b}{(A_a + A_b)/2}, \quad (10.17)$$

differential pointing \vec{d}

$$\vec{d} \propto \frac{\vec{b}_a - \vec{b}_b}{2}, \quad (10.18)$$

differential ellipticity

$$q \propto \frac{e_a - e_b}{2}, \quad (10.19)$$

and cross polarization. We use the “proportional to” notation above because authors differ on the normalization of the different quantities. The physical interpretation of these errors is straightforward. Each of the Q and U Stokes parameters that quantifies the polarization content of incident radiation is formed by differences of intensities between two orthogonal polarization states. A “differential gain” systematic error arises when the overall beam size, or the antenna gain, is different between any two orthogonal states.¹⁷ “Differential pointing” and “ellipticity” arise when there is a difference in the centroid of the beams and their ellipticities, respectively. It is not difficult to see that respectively these effects couple the temperature anisotropy, its gradient and its second derivative into the polarization measurement (see also Yadav et al. 2010). These errors are of the general category of “instrumental polarization,” in which unpolarized radiation becomes partially polarized by the instrument. It is a different category from cross polarization, also called “polarization rotation,” which acts only on polarized light, and its physical effect as far as the CMB is concerned is to mix E- and B-modes (see [Sect. 1.5](#)). All of the errors should be minimized through the design of the optics, or calibrated with an accuracy commensurate with the goal of the experiment.

Both instrumental polarization and polarization rotation occur because reflection and absorption, and hence transmission, of light depend on the polarization state of the incident radiation, on the angle of incidence, and, where relevant, on the materials making lenses or other optical elements (such as a half-wave plate, vacuum window, or reflectors). As an example, consider the effects of a standard ~ 0.05 -mm-thick vacuum window made of ~ 9 -cm-diameter polypropylene, which has an index of refraction close to 1.5. This will have high mm-wave transmission. Such a window, which is similar to the one used on the MAXIPOL polarimeter, is naturally bowed by few cm into the cryostat because of differential pressure. MAXIPOL had an array of 16 photometers with an intermediate focus near this window and the angle of incidence of rays for an edge photometer spanned 20 – 55° . Differential reflection between the two polarization states is about 1%, which generates a $1 \mu\text{K}$ spurious, scan-synchronous polarized

¹⁷Differential gain also arises when other factors in the system affect gain between two polarization states, for example, when two independent detectors that are sensitive to the two polarization states have different responsivities.

signal from the $\sim 100 \mu\text{K}$ *rms* CMB anisotropy. This spurious signal was of no consequence for the analysis of the MAXIPOL data, but is one example of an instrumental polarization that may be a contaminant for future, higher sensitivity experiments. A much larger signal of $\sim 12 \text{ mK}$ is generated by the 0.5% emissivity of the 250-K primary reflector.¹⁸ However, if that signal is stable, it can be removed in analysis as an overall offset. Of course, reflection from the asymmetric primary reflector, which is an off-axis section of a parabola, also polarizes unpolarized light, but this effect is very small, and, again, if the temperature is relatively constant, this overall polarized offset, or even if it is slowly drifting, can be removed in analysis.

As mentioned above, the systematic errors associated with the optics are either negligible compared to statistical errors in the measurement, or they need to be calibrated with an accuracy that makes them negligible. The faintness of the B-mode signal makes it susceptible to the various systematic errors, and this is our focus in the following discussion. In terms of absolute magnitude, the most challenging effects are differential gain and polarization rotation. Differential gain is challenging because it couples the temperature anisotropy directly into the polarization measurement.¹⁹ A residual differential gain of less than 0.1% is necessary if the leakage from temperature anisotropy is to be less than 10% of the B-mode predicted power spectrum with $r = 0.1$ (Zaldarriaga, 2006, private communication). Because polarization rotation can mix E- and B-modes, an uncorrected rotation of 0.3° of the incident polarization by the instrument gives rise to a spurious B-mode that is a factor of 10 below the B-mode signal from lensing. This is the level that should reasonably be targeted by an instrument that intends to *detect* the B-mode from lensing and not be limited by this systematic effect. A rotation of 1.3° (0.4°) gives rise to a spurious B-mode that is a factor of 10 below the cosmological inflationary signal with $r = 0.1$ (0.01) (Zaldarriaga, 2006, private communication).

The crossed-Dragone and the on-axis refractive-only systems appear to provide the lowest levels of the systematic errors discussed above. The EPIC team has quantified the expected level of instrumental polarization in their proposed crossed-Dragon telescope (Tran et al. 2010) and showed that the telescope nearly satisfies mission goals. Exceptions included differential gain, which is apparently caused by differential attenuation due to the finite conductivity of the aluminum surface, and differential rotation, which was up to 0.6° at the edge of the focal plane (Johnson, 2011, private communication). Therefore, both would need to be calibrated and corrected if the mission is to achieve its goals of setting a limit on $r = 0.01$. Imbriale et al. (2011) show measurements from QUIET's crossed-Dragone that demonstrate the absence of cross polarization at the center of the FOV.

A thorough analysis of beam-induced systematics for a refractive system was presented in conjunction with the release of data from BICEP (Takahashi et al. 2010). All measured systematic effects were smaller than benchmarks calculated for $r = 0.1$. Many measured effects were also acceptable for $r = 0.01$ with the two exceptions: “relative gain” and differential pointing.²⁰ The origin of the relative gain mismatch is likely the detection system and not BICEP's centered refractive optics. The observed differential pointing is an intriguing effect also observed by the QUAD experiment (Hinderks et al. 2009). The centroids of beams corresponding to two polarization-sensitive bolometers, which *do* share the same optical path, were offset relative to

¹⁸A 1.2 K nearly unpolarized signal is generated by the 0.5% emissivity. This signal is differentially polarized at the 1% level.

¹⁹The 2.7 K CMB monopole can also lead to a polarized signal through instrumental polarization. However, the magnitude of this signal is a constant across the observation, and since essentially all CMB polarimeters are differential, they are not sensitive to this overall offset.

²⁰Although Takahashi et al. (2010) define relative gain slightly differently than in (10.17), the underlying physics is the same.

each other. The teams speculated birefringence in the high-density polyethylene lenses, among other causes (Pryke, 2011, private communication). The effect was not of any consequence for the analysis of the experiments' data, and the BICEP team reports sufficient sensitivity to remove the effect if it had been necessary. If it persists, then future refractive telescopes searching for r values smaller than about 0.1 will also need to characterize the effect and subtract it during the analysis of the data.

Tran et al. (2008) compare the polarization performance of the cross and Gregorian systems. They show that in every respect the crossed-Dragone has superior polarization properties compared to the Gregorian system. For example, at a field point 3° away from the center of the focal plane of the Gregorian-Dragone system, the cross-polar response is -25 dB below the copolar response, whereas for a similar aperture crossed-Dragone the response is -50 dB. We note, however, that once the need arises to calibrate the effects induced by the optical system, the difference in performance between the Gregorian and crossed systems may become inconsequential.

3 Large Ground-Based Telescopes, ACT and SPT

With 6- and 10-m primary reflectors, the ACT (Fowler et al. 2007; Swetz et al. 2011) and SPT (Carlstrom et al. 2011) (🔗 Fig. 10-7) are currently the largest CMB survey telescopes providing the highest resolution. Reviewing them in more detail is instructive because the designs incorporate the state of optics knowledge for ground-based instruments as of mid-2000.

There are a number of reasons to probe the CMB with high resolution. As can be seen in 🔗 Fig. 10-2, arcminute resolution, and hence a window function that extends to $\ell > 5,000$, is required to determine the contribution of point sources and other secondary sources of anisotropy (not shown) to the damping tail of the primary CMB anisotropy. It is also advantageous to have minimal change in the beam over $\ell < 3,000$, the region where the primary anisotropy dominates. Even with its 6-m primary, the ACT beam suppresses the fluctuations by a factor of 3 in power by $\ell = 6,000$. Lastly, one can discover galaxy clusters via the thermal Sunyaev-Zel'dovich (SZ) effect (Sunyaev and Zeldovich 1980) – the inverse Compton scattering of CMB photons with hot electrons in ionized gas – which requires ~ 1 arcminute resolution in bands near 150 GHz.

The design of the ACT and SPT optical systems are shown in 🔗 Figs. 10-8 and 🔗 10-9. They are both based on off-axis Gregorian telescopes. The SPT secondary mirror is maintained at cryogenic temperatures and an additional lens focuses the light onto the telecentric focal plane. In ACT, the secondary mirror is maintained at ambient temperature. At the entrance to the cryogenic receiver, the focal surface provided by the the Gregorian telescope is split into three distinct light trains, each dedicated to a single frequency band. An off-axis Gregorian telescope was chosen for both systems because of the guiding principles listed in 🔗 Table 10-2, which are also discussed in 🔗 Sect. 2.5. In addition, it is easier to implement co-moving ground shields and to minimize spillover to the surroundings on the sides of an off-axis Gregorian.

3.1 Detailed Optical System Comparison

The SPT is a standard Gregorian-Dragone design (D1, in the language of 🔗 Sect. 2.5) (Padin et al. 2008). The SPT Gregorian configuration and the primary diameter, parabolic shape, and



■ Fig. 10-7

Left: The ACT 6-m telescope in northern Chile. The telescope is inside the 13-m-tall ground screen. The secondary is just visible near the center of the ground screen. *Right:* The SPT 10-m telescope at the South Pole. Most of the circular primary reflector is visible, while the cryogenic secondary reflector and receiver are housed inside the white structure below and to the right of the primary (Photos courtesy of ACT and SPT Collaborations)

■ Table 10-2

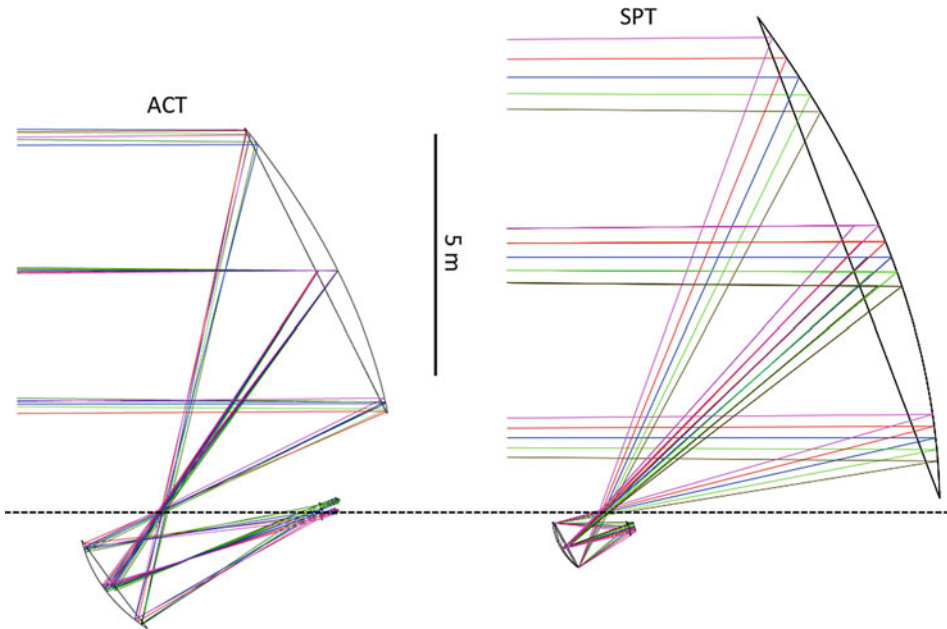
Guiding principles of the ACT and SPT designs

• Clear aperture (off-axis optics) to minimize scattering and blockage
• Fast primary focus to keep the telescope compact and enable fast scanning
• Large (FOV $\sim 1^\circ$) and fast ($F \sim 1$) diffraction-limited focal plane
• Space for structure and a cryogenic receiver near Gregorian focus

surface accuracy ($20 \mu\text{m rms}$) were designed to accomplish a wide range of millimeter and sub-millimeter science goals. The initial ACT design started from a standard Gregorian-Dragone (D1); however, the conic constants of both ACT reflectors were then numerically optimized to maximize the DLFOV area by minimizing the transverse ray aberration across the focal plane (Fowler et al. 2007).²¹ This leads to an aplanatic-like solution in which the primary reflector becomes elliptical instead of parabolic. The ACT reflector optimization converged to a 150-GHz DLFOV at the Gregorian focus of $\sim 370 \text{ cm}^2 \text{ sr}$ with near zero tilt of the secondary reflector axis, which led to the decision to align the ACT primary and secondary reflector axes to simplify the system alignment (► Fig. 10-8).

The secondary reflectors of ACT and SPT meet substantially different design requirements. The SPT secondary was designed to meet the science goals of the first generation camera and

²¹We note that this optimization approach in which the conic constants of both reflectors are simultaneously optimized across a flat focal plane is different from the optimization discussed in Hanany and Marrone (2002), where the focal plane shape and position were optimized to provide the largest DLFOV.



■ Fig. 10-8

Ray trace diagram of the ACT and SPT optical systems. ▶ [Figure 10-9](#) shows an enlarged view of the parts of the systems that are maintained at cryogenic temperatures. The distribution of rays across the primary reflectors is indicative of the proximity of the cold stop to an image of the primary reflector. Having a stop at an image of the primary in ACT enables more uniform illumination of the primary (▶ [Table 10-3](#)). The *dashed line* shows the conic axis of the primary reflector of both telescopes, which is also the conic axis of the secondary reflector for ACT. The SPT secondary reflector conic axis is tilted by 25.3° to minimize aberrations and cross polarization and maximize the FOV, which also moves the Gregorian focus below the primary axis, providing space for the receiver. The lower edge of the ACT primary is further off-axis than SPT, which provides space for the receiver without tilting the secondary reflector (SPT ray trace courtesy of N. Halverson)

has a short focal ratio ($F_G \approx 1.3$) to couple to a close-packed feedhorn array at the Gregorian focus. The entire secondary is cooled to ~ 10 K inside a vacuum vessel and is used at the aperture stop of the system. Because this stop is not quite at an image of the primary, different detectors across the focal plane illuminate different sections of the primary reflector. This is one reason that only ~ 7.5 m of the 10 m SPT primary reflector is illuminated by each feedhorn (▶ [Table 10-3](#), ▶ [Fig. 10-8](#)). To ensure that the secondary reflector could be aligned accurately inside the cryogenic receiver, it was machined as a single piece of metal, which limited the size to ~ 1 m diameter – the largest diameter that could easily be machined as a single element.

The ACT secondary reflector focal ratio ($F_G \approx 2.5$) was selected as a balance between smaller reimaging optics and minimizing beam expansion because of the need to closely pack neighboring stacks of optical elements at multiple cryogenic stages (including a vacuum window, filters at 300, 40, and 4 K, and a lens at 4 K) near the Gregorian focus. The secondary diameter was chosen to be ~ 2 m as a balance between taking advantage of the increase in diffraction-limited throughput that a larger secondary provides and minimizing the mass far from the telescope

■ Table 10-3

Comparison of some ACT and SPT optical parameters. The top section gives the main telescope parameters; the middle section gives the properties of the receivers deployed prior to 2012; and the bottom section is for the polarization-sensitive receivers. D_p , D_i , and D_s are the diameters of the primary reflector, the illumination of the primary, and the secondary reflector, respectively. F_G , F_{temp} , and F_{pol} refer to the approximate focal ratio at the Gregorian, temperature receiver, and polarization receiver foci, respectively. $F\lambda_{150}$ refers to the approximate 150-GHz detector spacing or feedhorn aperture for each receiver. $A\Omega_{R_{\text{temp}}}$ and $A\Omega_{R_{\text{pol}}}$ refer to the approximate throughput of the “temperature receivers” (for all three arrays) and “polarization receivers” for each telescope at the detector focus, and $A\Omega_{D_{\text{eff}}}$ is an estimate for the relative effective throughput of the different detector array technologies following the prescription of Griffin et al. (2002). For the temperature receivers, the minimum 150-GHz Strehl ratios and the measured 150-GHz beam full-width-half-maximum, FWHM_{150} , are also provided (Swetz et al. 2011; Schaffer et al. 2011)

	ACT	SPT
D_p (m)	6	10
D_i (m)	5.6	7.5
D_s (m)	2	1
F_G	2.5	1.3
Temperature receivers		
F_{temp}	0.9	1.3
$F\lambda_{150}$	0.5	1.7
$A\Omega_{R_{\text{temp}}}$ ($\text{cm}^2 \text{sr}$)	40	105
$A\Omega_{D_{\text{eff}}}$ (relative)	~2.5	1
Min. Strehl ₁₅₀	0.97	0.89
FWHM_{150} (arcmin)	1.37	1.15
Polarization receivers		
F_{pol}	1.4	1.3
$F\lambda_{150}$	1.4	1.6
$A\Omega_{R_{\text{pol}}}$ ($\text{cm}^2 \text{sr}$)	180	140

center-of-mass (to enable fast scanning) and cost. The combination of the aplanatic design, a larger secondary, and a higher F result in the ACT having substantially greater diffraction-limited throughput at the Gregorian focus, than the SPT; however, the ACT design requires an additional set of reimaging optics because the secondary reflector cannot easily be incorporated into a cryogenic receiver (like the SPT secondary). This results in increased thermal emission from the ACT secondary and constrains use of the ACT Gregorian DLFOV because of the difficulty of building a compact, low-loss vacuum window as large as the DLFOV (~0.7 m diameter).

One characteristic of optimized aplanatic designs, like ACT, is that the Gregorian focal plane becomes more perpendicular to the conic axis of the reflectors, which is an advantage for on-axis systems, but results in a less image-space telecentric focal plane for off-axis systems. For example, in the ACT design, the angle of incidence of the chief ray at the center of the Gregorian focal plane is 18.7° . A telecentric focal plane was not a requirement for the ACT receiver, and at higher F numbers, this “focal plane tilt” is reduced, which led to an acceptable level of residual $5\text{--}8^\circ$ tilts

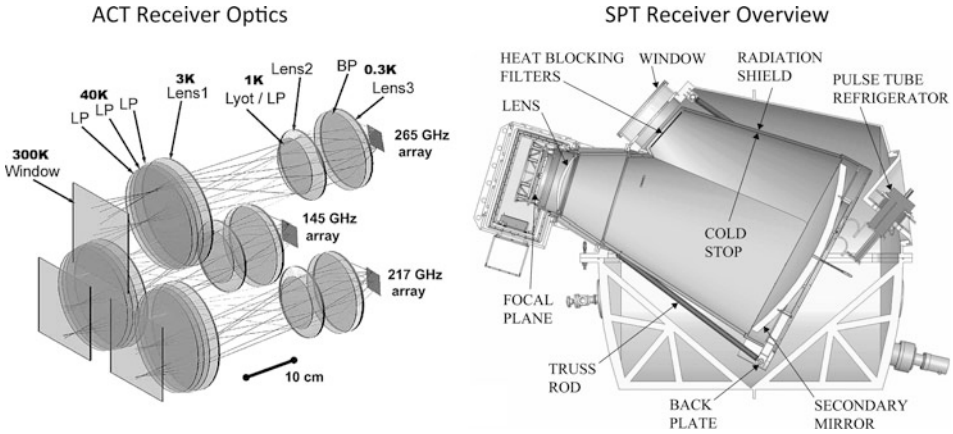
at the filled-focal-plane arrays used in the ACT receiver. However, coupling to a tilted focal plane does require custom reimaging optics, which generally prevents installing receivers from other telescopes without reconfiguring the receiver optics to match the ACT focus. In addition, when considering designs for future instruments that may require a different F , it is straightforward to analytically calculate the new secondary parameters to match a parabolic primary, like SPT. It has previously been stated that in an aplanatic Gregorian design, the range of secondary focal lengths is limited (Padin et al. 2008); however, the ACT design demonstrated that numerical optimization techniques can be used to adjust the focal length of the secondary with the design of an ellipsoidal primary. As the secondary F is increased (decreased), the DLFOV throughput increases (decreases), which is the same behavior as a classical Gregorian with a parabolic primary. The limits of changing F in an optimized aplanatic Gregorian design (ACT) relative to a classical Gregorian (SPT) are not clear and warrant further study. Because of the focal plane tilt and numerical optimization requirements of changing the secondary F , in [Table 10-4](#), we characterize the SPT design as “easily reconfigurable” but not the ACT design.

Both the ACT and SPT teams began observing with multifrequency receivers to measure the CMB and to search for galaxy clusters via the SZ effect in 2008 and are planning to deploy new polarization-sensitive receivers in 2012–2013. Here, we briefly compare the first-generation receiver optics ([Fig. 10-9](#)) and detector coupling, then discuss the planned upgrades. The SPT receiver has two sections that can be operated independently, which is beneficial for testing both systems and for upgrades that will use the same secondary reflector. The first section includes the vacuum window, thermal blocking filters, cryogenic secondary reflector, and the majority of the cold stop. The second section contains a lens, band-defining filters, and the detector array, which includes detectors at ~ 100 , 150, and 220 GHz. Some advantages of this design include reduced emission from the secondary reflector, a small-aperture vacuum window at the $F \approx 1$ primary focus, and a large stop surface, which minimizes diffraction at the stop. The SPT detectors are coupled to the optics via a flat array of conical feedhorns, so a single high-density polyethylene ($n \approx 1.5$) lens is used to slightly speed up the focus and to improve the coupling to the feedhorns by making the focal plane more telecentric. The feedhorn-coupled array includes 966 feeds with 4.5-mm apertures, resulting in $\sim 1.7\text{-}F\lambda$ apertures at 150 GHz.

The ACT temperature receiver includes three independent optical paths that each operate at a different frequency band: 148, 218, and 277 GHz. Each optical path has an independent vacuum window, filters, three silicon ($n \approx 3.4$) lenses, and detector array, which makes defining the bandwidth and antireflection coating the optical elements relatively easy compared to

Table 10-4
Comparison of some ACT and SPT design features

	ACT	SPT
Primary shape	Ellipsoid	Paraboloid
Easily reconfigurable	No	Yes
Stop type	Primary image	Secondary reflector
Cold stop temperature	1 K	10 K
Refractive optics	Three silicon lenses per array	HDPE lens
Temp. array coupling	Filled focal plane	Conical feedhorns
Pol. array coupling	Corrugated feedhorns	Corrugated and profiled feedhorns



■ Fig. 10-9

Left: The ACT receiver optics include three independent optical paths, each for a different frequency band, with its own vacuum window, filters, and set of three silicon lenses (Fowler et al. 2007). **Right:** An overview of the SPT receiver, which includes the cryogenic secondary reflector and a single HDPE lens (Padin et al. 2008). The secondary mirror is the aperture stop of the system; spill-over past the secondary is intercepted by cold surfaces

receivers that use common optical elements for multiple frequency bands. The first lens in each optical path creates an image of the primary reflector, which is used as the system stop and allows illumination of $>90\%$ of the primary reflector (► Table 10-3). The stop surface is cooled to 1 K, which is required to minimize the background optical loading on the filled detector arrays from spillover onto the stop. The following pairs of lenses create a fast focus ($F \approx 0.9$) onto the three filled detector arrays, each comprised of 1,024 square bolometric detectors with 1.1 mm pitch (Niemack et al. 2008), or roughly $0.5F\lambda$ at 150 GHz.

► Section 2.3 provides an overview of the trade-offs between filled focal plane arrays (used in ACT) and feedhorn-coupled arrays (used in SPT). The close packing of the detectors in ACT led to having ~ 3 times more detectors, despite the ACT receiver having less than half the optical throughput of the SPT receiver (► Table 10-3). We estimate that the ACT 150-GHz filled array could have ~ 2.5 higher mapping speed than a feedhorn array similar to the SPT design filling the same FOV (listed as $\Omega_{D_{\text{eff}}}$ in ► Table 10-3). Scaling the mapping speed ratio by the total instrument throughput, $A\Omega_{R_{\text{temp}}}$, provides an estimate of the relative mapping speeds of the two instruments of $MS_{\text{SPT}}/MS_{\text{ACT}} \approx 1.05$. This estimate includes assumptions about the instrument optical loading conditions, does not include detector noise, and assumes that 150-GHz detectors fill the same fraction of $A\Omega_{R_{\text{temp}}}$ of both instruments. Based on the mapping speed estimate, the increased throughput of the filled detector arrays on ACT largely compensates for the smaller $A\Omega_{R_{\text{temp}}}$. This suggests that filled detector arrays hold promise for maximizing the mapping speed of future instruments; however, significant development is needed to scale the readout and fabrication of filled arrays for systems with larger $A\Omega$. In addition to simplifying instrument requirements as discussed above, feedhorns (and other beam-forming detector coupling techniques) have advantages in terms of minimizing systematic effects in polarization measurements.

3.2 ACTPol and SPTPol

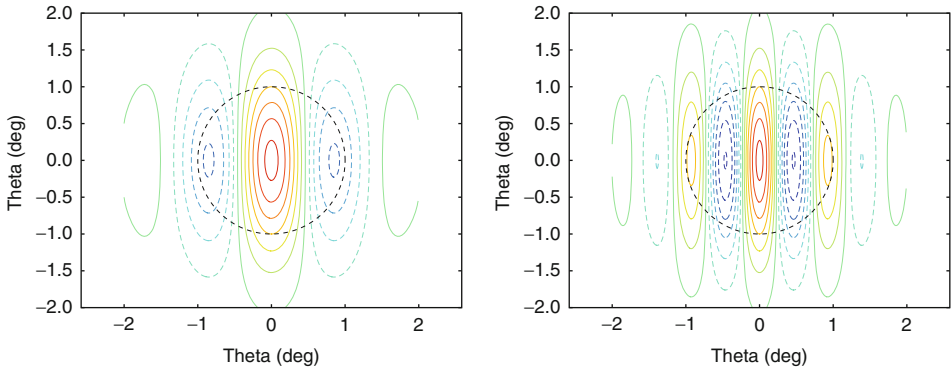
The ACT and SPT polarization-sensitive receivers, henceforth ACTPol (Niemack et al. 2010) and SPTPol (McMahon et al. 2009), both use corrugated feedhorns (as well as 90-GHz profiled feedhorns in the case of SPTPol) to couple to the detector arrays. This is done because of the excellent polarization properties of corrugated/profiled feedhorns. SPTPol uses a nearly identical optics layout to the SPT layout, but the used FOV area has been increased by $\sim 33\%$ by surrounding the central 150-GHz detectors with individual 90-GHz feedhorns. ACTPol uses a similar optics concept to ACT (three independent optics paths with three silicon lenses each); however, the diameter of most of the optical components is roughly two times larger, and the detector arrays are circular instead of square, which leads to a factor of 4.5 increase in throughput at the final focus (► [Table 10-3](#)). Unlike the original ACT lenses, the orientation of the ACTPol lenses is optimized to create image-space telecentric focal planes to couple to the flat feedhorn arrays.

The ACT design approach included maximizing the throughput of the DLFOV at the Gregorian focus, but the DLFOV does not necessarily limit the usable FOV. For example, the edges of the ACTPol lenses that are furthest from the boresight are placed outside the Gregorian DLFOV; however, the lenses improve the image quality in this region to be strongly diffraction limited (Strehl ratios > 0.9). Adding a tertiary reflector to a design like SPT is another approach for increasing the DLFOV. Thus far, the throughput of the telescope receivers has generally been designed to match the requirements of the superconducting detector arrays that were feasible to manufacture and read out at that time. As superconducting detector array technologies continue to increase in size and decrease in cost, higher throughput optical systems will be required to illuminate them. These examples suggest that modified secondary, tertiary, or reimaging optics may allow substantial increases in the useable FOV for these telescopes in the future.

4 Interferometers

Interferometers are a natural choice for measuring the anisotropy in the CMB. The correlation of the outputs from two antennas, called a visibility, is just a Fourier component of the product of the sky and the response of a single antenna. The single antenna response is called the primary beam and sets the interferometer's field of view. It should not be confused with the response of the pair of telescopes (or more generally, of an array of telescopes) to a point source, called the synthesized beam. Thus, the visibility is directly relatable to the CMB power spectrum. In the following, we imagine that the region of sky we cover is small enough that we may consider it flat so that the expansion of the temperature field in $Y_{\ell m}$ spherical harmonics can be replaced by Fourier modes.

Consider the case of two identical antennas (or telescopes) of diameter D that are separated by distance L . A particular configuration of the antennas, set by the spatial separation L is called a 'baseline'. When $L = 2D$ the antennas are in a 'compact configuration'. The instantaneous beam pattern on the sky would resemble that of a double slit pattern for two wide slits. A representation is shown in ► [Fig. 10-10](#). A visibility is the instrument response of the sky times this instantaneous beam. As the telescopes are moved apart, the envelope of the pattern remains that of the beam pattern of a single telescope although the number of fringes inside the envelope increases. With multiple measurements with baselines of different lengths and orientations, one



■ Fig. 10-10

Instantaneous interferometer beams. Consider a 30-cm-diameter aperture with a Gaussian profile with $\sigma_r = 6.3$ cm (► Sect. 1.2) and $\lambda = 1$ cm. The resulting beam has $\theta_{1/2} = 2^\circ$. This is indicated by the *dashed circular lines* in the figures. If two such apertures are placed so that the center-to-center separation is 30 cm (30λ), then one obtains the instantaneous beam pattern shown on the *left*. The *dashed lines* indicate negative lobes and the *solid lines* positive lobes. The output of the interferometer is then the integral of the product of this beam pattern and the sky. One can see that only spatial fluctuations that resemble this pattern will give a nonzero output. If the center-to-center separation is increased to 60λ , then one obtains the instantaneous pattern on the *right*. Note that the extent of the pattern is still determined by $\theta_{1/2}$ from one antenna

may fill out the “U-V” (or loosely Fourier) plane with visibilities. To compute the power spectrum, one averages the variance over annuli in the U-V plane. To make a map of the sky, one then transforms the visibility map to real space. There is no reason to constrain oneself to the envelope of the beam, many images like the one in ► Fig. 10-10 can be mosaicked together to probe spatial wavelengths longer than that of the beam size and to increase the resolution of the visibility spacing.

In some sense, interferometers are the opposite of more conventional mapping schemes. With real-space maps, the size of the beam sets the resolution, and the scan size of the instrument sets the size of the field one observes. With interferometers, the separation of the telescopes determines the resolution and the transform of the antenna illumination pattern sets the field size (before any mosaicking). With real-space methods, a map is made and the power spectrum is obtained from its Fourier transform. With interferometers, the power spectrum is fairly directly measured from the visibilities. The real-space maps can be obtained from the transform of the visibilities, though they are rarely used for CMB science.

Interferometers have been used to measure the CMB anisotropy since the mid-1980s (e.g., Knöke et al. 1984 and reviewed in Partridge 1995), although most of the early efforts were aimed at arc-minute angular scales or smaller because they used the VLA. The first interferometric observation of the primary CMB was made with a dedicated two-element correlation receiver (Timbie and Wilkinson 1988). Results from CAT, when combined with results at lower ℓ , gave compelling evidence for the existence of the first acoustic peak in the CMB (O’Sullivan et al. 1995; Scott et al. 1996; Baker et al. 1999). Anisotropy measurements by the 13 antennas of the Cosmic Background Imager interferometer at 31 GHz (Pearson et al. 2003),

together with results from ACBAR (Kuo et al. 2004) and WMAP, helped break cosmic parameter degeneracies. The DASI interferometer made the first measurements of the polarization of the CMB (Kovac et al. 2002).

Interferometers have a number of advantages over real-space methods. The spatial filtering of a visibility makes it insensitive to scales much larger (or smaller) than the fringe spacing, so interferometers filter out almost all atmospheric fluctuations during the correlation. They can be set up to measure fine angular resolution easily, simply by increasing the baseline length. As one adds elements, the number of baselines grows as the square of the number of antennas. The relative response to different spatial wavelengths (analogous to the beam of a single-dish telescope) is set by the easily measured separation of antennas. Because of the intrinsic atmospheric filtering, they do not have to scan rapidly. The primary disadvantage is that the cost of correlation in a classical interferometer grows as the number of dishes squared (though see, for example, the Fast Fourier Transform Telescope (Tegmark and Zaldarriaga 2009, 2010)). For sparsely sampled arrays interferometers are slower at mapping the sky; a separate cryostat is required for each receiver; and the components are expensive. At 150 GHz coherent receivers are not yet a “commodity” and the mechanical tolerances are tight. With the advent of arrays of thousands of bolometers and of order 100-element polarization-sensitive coherent receivers (Gaier et al. 2003), *classically configured* interferometers have lost much of their appeal for measuring the CMB.

However, the quest for primordial gravitational waves and the advantages of interferometers have driven the invention of new designs that go well beyond the classic configuration. There is now an international effort called QUBIC (The Qubic Collaboration et al. 2011; Charlassier 2008; Timbie et al. 2006) to use arrays of bolometers in a novel optical configuration to make an interferometer capable of detecting the polarization B-modes. The instrument observes the sky through an array of feeds whose signals are then interferometrically combined on two arrays of $\sim 1,000$ bolometers, with one array per polarization. As opposed to multiplying the signals from a pair of antennas, in QUBIC, the interfering electric fields are summed and squared by the bolometers. It is a modern version of the Fizeau-style adding configuration.

5 The CMB Satellites

There have been four satellites with instruments dedicated to measuring the CMB anisotropy: Relikt (Strukov and Skulachev 1984), the COsmic Background Explorer (COBE) (Boggess et al. 1992), the Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al. 2003), and Planck (Tauber et al. 2010b). The frequency coverage, sensitivity, and resolution are given in [Table 10-5](#). There is a marked improvement over time.

The great benefit of a satellite is the ability to make all-sky maps from a very stable platform. The stability of space allows one to understand the instrument, especially the noise and systematic effects, in detail. An ideal anisotropy map would be fully described by simply a temperature and uncertainty for each pixel on the sky with an overall offset removed. In reality, all maps have some degree of correlation that must be accounted for in the most demanding analyses. The source of the correlation could be due to nonideal aspects of the instrument's noise (e.g., “ $1/f$ ” noise), remnants of glitch removal, contamination through the sidelobes, an imperfect differential measurement (for COBE and WMAP), or asymmetric optics. Multiple techniques have been developed to account for these correlations.

An important factor in making high-fidelity maps is cross-linking. In the limit of a perfectly stable instrument, cross-linking is not necessary, but in reality the gain and offsets of all detectors vary over time. From the point of view of one pixel, a well cross-linked map would have scan lines running through in all different directions connecting the pixel in question to those around it. Ideally the cross-linking takes place on multiple timescales. A set of such measurements for each pixel over the full sky provides a strong spatio-temporal filter that allows for the separation of instrumental effects such as varying gains and offsets from the true underlying signals.

Cross-linking has an advantage for the optics as well. Because of the premium on size and mass for a satellite, the focal planes are packed to the hilt. As a result, the beam profiles are not symmetric let alone Gaussian. Perfect cross-linking has the effect of producing an effectively symmetric beam profile (with a coarser resolution that depends on the inherent asymmetry), thereby simplifying the data analysis. As with correlations, the most demanding analyses must take the remaining asymmetries into account (Hinshaw et al. 2007; Hanson et al. 2010).

5.1 Relikt

Relikt was the first space-based anisotropy satellite (Strukov and Skulachev 1984). It was part of the Soviet space program and was launched on the Prognoz-9 satellite in 1983, roughly 6 years before the launch of NASA's COBE. The microwave radiometer, shown in Fig. 10-11, was one of a number of instruments on the satellite. Its three objectives were

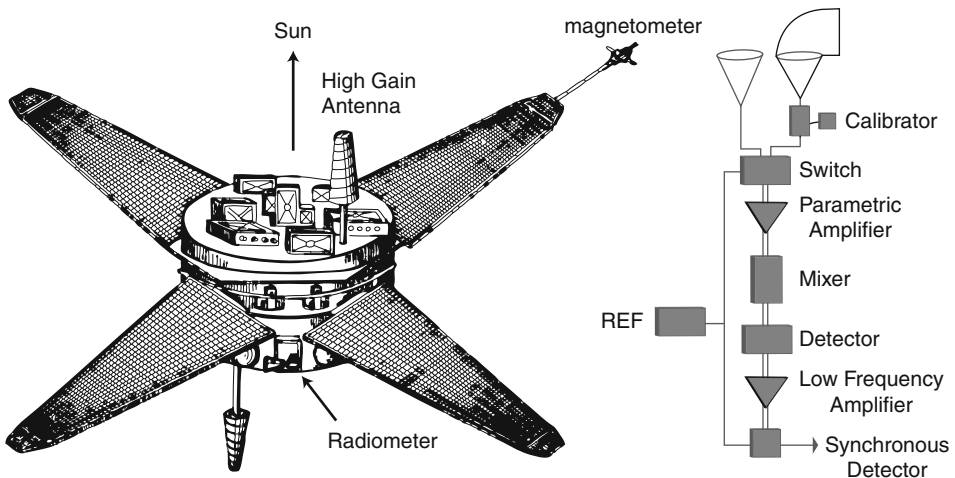


Fig. 10-11


A line drawing of the Relikt satellite from Strukov and Skulachev (1986). The Sun is toward the top of the page. The radiometer is the small box in the center between the bottom two solar panels. The reference feed points toward the bottom of the page and the scanning feed points toward the left. The instrument package weighed only 30 kg



(Strukov and Skulachev 1986) “(1) to determine the angular distribution of the relic radiation and (in the case of the discovery of anisotropy) to estimate the mean density of matter in the universe; (2) to determine the distribution of faint extended radio sources on the celestial sphere; and (3) to refine the velocity-vector parameters of the observer’s motion with respect to the reference frame of the relic radiation.” Outside of detecting the primary anisotropy, these goals were achieved.

Relikt was a differential instrument although was not symmetric. A reference-corrugated feed with $\theta_{1/2} = 10^\circ$ pointed in the antisolar direction. A second scanning feed with $\theta_{1/2} = 5.8^\circ$ was aimed 90° to the reference direction and scanned the sky as the satellite rotated. The satellite was in a highly elongated orbit, with a 26.7-day period. The rotation period was 113 s. In certain parts of the orbit the Earth was observed as the beam scanned over it. After averaging data for a week, the reference beam was stepped by 7° in the ecliptic. Thus, all scans overlapped at the ecliptic poles. Nearly the full sky was observed in the 6 months the satellite was operational (Klypin et al. 1992).

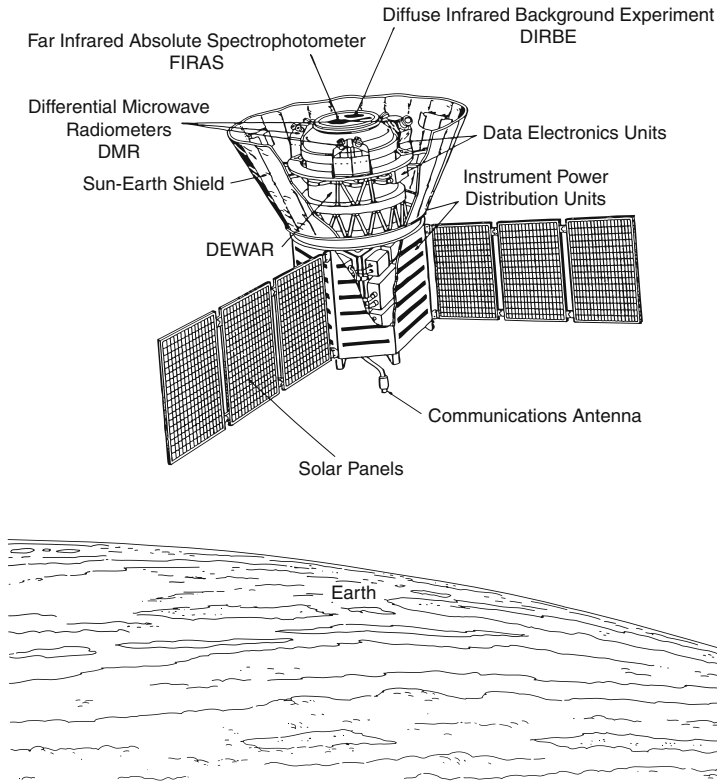
The scanning feed had a corrugated cone feeding an offset parabola to direct the radiation perpendicular to the symmetry axis of the base of the cone. Before launch the beam pattern was mapped (Strukov and Skulachev 1984) to -80 dB of the peak and a broad sidelobe at the -40 dB level for angles 30 – 60° was measured. As a result, during the data analysis, roughly half the data were cut due to possible contamination by emission from the Earth and Moon into the scanning beam sidelobes (Klypin et al. 1992). Nevertheless, at the time, they produced the best measurement of the dipole (Strukov et al. 1987), one of their primary goals, and they placed limits on the large angular scale anisotropy (Strukov et al. 1988) that were not improved upon until FIRS (Meyer et al. 1991) and COBE (Smoot et al. 1991).

5.2 COBE

The Differential Microwave Radiometer (DMR) instrument (Smoot et al. 1990) was one of three aboard the COBE satellite, shown in  Fig. 10-12. The other two were the Far Infrared Absolute Spectrophotometer (FIRAS) which measured the absolute temperature of the CMB with FIRAS ($T_{\text{CMB}} = 2.725 \pm 0.001$ K) (Fixsen and Mather 2002) and the Diffuse Infrared Background Experiment (DIRBE) which mapped the IR sky and detected the cosmic infrared background.

The COBE satellite was launched in 1989 into a high inclination polar orbit. The spin axis of the spacecraft always pointed away from the Earth and at roughly 90° from the Sun. At each of the three frequencies given in  Table 10-5, there were two receivers like the one shown in  Fig. 10-13. The radiometers were situated in the satellite so that the feeds observed $\pm 30^\circ$ from the spin axis. COBE’s orbit and scan pattern were a marked improvement over that of Relikt’s. Detailed attention was paid to possible contamination by emission from the Sun, Earth, and Moon, and large cuts of the data were not required (Kogut et al. 1992). The most notable systematic error was due to the affect of the Earth’s magnetic field on the Dicke switches.

The DMR optics are especially simple. In each of the six receivers, the sky is viewed through two corrugated feeds with $\theta_{1/2} = 7^\circ$ and separated by 60° . There are, though, only five pairs of feeds. At 31.5 GHz, the two output polarizations of feed pair are sent to two receivers. At the other frequencies, a single polarization from each feed pair is sent to a receiver. The beam patterns were measured before flight to roughly -90 dB from the peak (Toral et al. 1989).

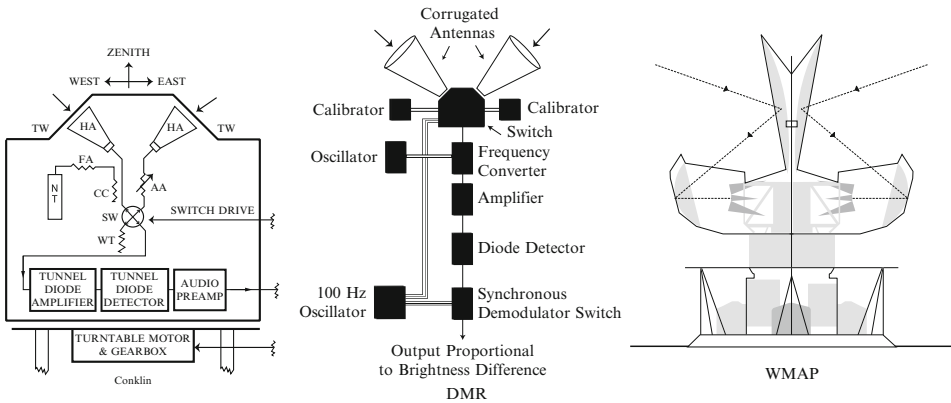


■ Fig. 10-12

A line drawing of the COBE satellite (Boggess et al. 1992). Radiometers for the three DMR bands are shown. The 31.5-GHz receiver (*back right*) has a single feed. The two separate radiometers in this band look at the two orthogonal polarizations from one feed. The Sun would be, for example, off to the left in this figure, thus illuminating the solar panels. The instruments always look away and are shielded from the Earth. For scale, the diameter with the deployed solar panels is 8.5 m. The mass of the DMR was 154 kg

The input to the receiver chain is Dicke-switched between the feed outputs at 100 Hz. The 31.5-GHz receiver operated at 300 K; the other two bands were at 140 K. The combined noise level for the three bands was 30, 11, and 16 $\text{mK}\cdot\text{s}^{1/2}$, respectively.

The DMR instrument was much different in layout than the one on Relikt. As shown in ● Fig. 10-13, it took advantage of symmetry. The first manifestly symmetric anisotropy instrument was designed by Edward Conklin (1969a, b) with the goal of measuring the CMB dipole from White Mountain, CA. While there is no apparent evolutionary connection between Conklin, COBE, and WMAP, the strong appeal of symmetry for making a differential measurement guided the designs of all these instruments. As we show below, most modern anisotropy instruments, including the Planck satellite, are not symmetric. In some cases, this is driven by the use of bolometers; in others by the fact that the receivers are dual polarized and thus intrinsically differential.



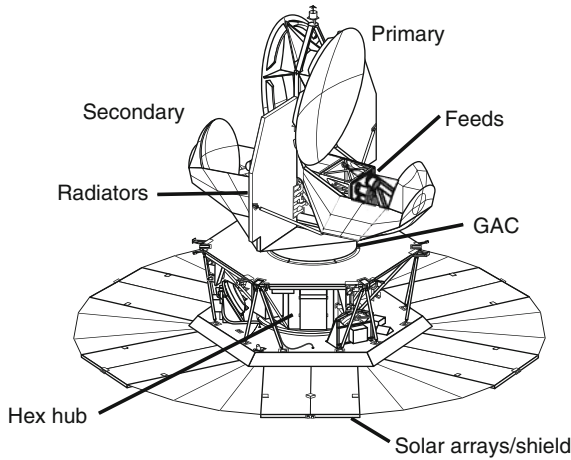
■ Fig. 10-13

The evolution of the manifestly differential CMB radiometer over 30 years. The *left* shows Conklin's radiometer for measuring the CMB dipole. The *central* picture shows the DMR aboard COBE. The angle between the feeds is 60° . The *right-most* figure shows WMAP. (Adapted from Peebles et al. (2009))

5.3 WMAP

As shown in ► Figs. 10-13 and ► 10-14, WMAP has similarities to a classic feed plus telescope design. However, unlike the classic system, two telescopes are combined in a back-to-back configuration. When designing a CMB satellite telescope, there are a number of factors that must be considered:



1. It is desirable to use an “offset” design so that the support structure for the secondary does not scatter radiation.
2. In space, one wants to make optimal use of the rocket shroud size. For a given useable focal plane area, the Gregorian family with its focus between the primary and secondary is especially compact (Brown and Prata 1994). For WMAP, multiple designs were considered including the offset Cassegrain, single reflector systems, and three reflector systems, but the two-mirror Gregorian was optimal.
3. One wants to get as many beams on the sky with as wide a frequency coverage as the technology will permit. In other words, one wants a large DLFOV. To minimize aberrations and maintain a DLFOV, the higher frequency feeds are placed near the center of the focal plane and the lower frequency feeds on the outside. Asymmetric beam profiles are acceptable as their effects may be incorporated in the data analysis. For WMAP, the scan strategy (Bennett et al. 2003) has the benefit of symmetrizing the beam profiles.
4. At least two types of modeling code are needed. A fast parametric code is useful for trying many designs. However, the full response must be computed using physical optics in which the field is solved for at each surface. WMAP used the Diffraction Analysis of a Dual Reflector Antenna code (Rahmat-Samii et al. 1995) which proved to be sufficiently accurate.
5. The precise reflector shape is chosen to minimize aberrations over as large as possible an area. WMAP started as a classic Gregorian following the Mizuguchi-Dragone condition (see ► Sect. 2.5 and footnote 15). Then the surface was shaped using proprietary surface-shaping software (Rahmat-Samii et al. 1995).



■ Fig. 10-14

Outline of the WMAP satellite. The thermal straps at the base of the large radiators provide the passive cooling for the first stage of the amplifier chain. The overall height is 3.6 m, the mass is 830 kg, and the diameter of the large disk on the *bottom* is 5.1 m. Six solar arrays on the *bottom* of this disk supply 400 W to power the spacecraft and instrument. Thermal blanketing between the hex hub and thermal link provided by the gamma alumina cylinder (GAC), and between the GAC and radiators, shield the instrument from thermal radiation from the support electronics and attitude control systems

6. The feeds must not be able to view each other or couple to each other. This means that low-frequency feeds are shortened or profiled and that high-frequency feeds are lengthened with extra corrugations.
7. One must be able to account for all of the solid angle of the beam in intensity and polarization. The optics are designed with the full 4π coverage in mind. Not only are the Earth, Moon, and Sun bright objects in the sidelobes, but emission from the galaxy must be considered. For WMAP, a specialized test range was built to make sure that, by measurement, one could limit the Sun as a source of spurious signal to $<1 \mu\text{K}$ level in all bands. This requires knowing the beam profiles down to roughly -45 dBi (gain above isotropic) or -105 dB from the W-band peak. It was found that over much of the sky, the measured profiles differ from the predictions at the -50 dB level due to scattering off of the feedhorns and the structure that holds them that were not part of the model. During the early part of the mission, the Moon was used as a source in the sidelobes to verify in part the ground-based measurements and models.
8. After launch and before commanding the satellite attitude, the optics have a chance of directly viewing the Sun. Thus, the surfaces must be roughened so that the Sun is not focussed on the feeds. The roughening must not increase the microwave emissivity. In addition, the surface must be emissive at infrared wavelengths so that it can radiatively cool. This is accomplished by evaporatively coating them with a mixture of silicon monoxide and silicon dioxide.
9. There is a premium on mass and therefore the reflectors are almost always made of lightweight composite material.


As a result of these considerations, the WMAP team settled on a design (Page et al. 2003; Barnes et al. 2002) shown in  Fig. 10-14. The primary reflectors are 1.4×1.6 m. The secondaries are roughly a meter across, though most of the surface simply acts as a shield to prevent the feeds from directly viewing the Galaxy. Each telescope focusses radiation onto ten dual-polarization scalar feeds. These are shown as triangles in the right panel of  Fig. 10-13. The primary optical axes of the two telescopes are separated by 141° to allow differential measurements over large angles on a fast timescale. The feed centers occupy a $18 \text{ cm} \times 20 \text{ cm}$ region in the focal plane, corresponding to a $4^\circ \times 4.5^\circ$ array on the sky.

At the base of each feed is an orthomode transducer (OMT) that sends the two polarizations supported by the feed to separate receiver chains. The microwave plumbing is such that a single receiver chain (half of a “differencing assembly,” Jarosik et al. (2003)) differences electric fields with two nearly parallel linear polarization vectors, one from each telescope.

Because of the large focal plane, the beams are not symmetric nor are they Gaussian. In addition, as anticipated, cooldown distortions of the primary reflectors distort the W-band and V-band beam shapes.

As noted above, precise knowledge of the beams is essential for accurately computing the CMB angular spectrum. For WMAP, one of the most CPU-intensive aspects of the data analysis was modeling the beams. The goal was to find an antenna pattern that matched the in-flight measurements of Jupiter using each of the two telescopes, the sidelobe measurements from the Moon, and preflight ground-based measurements. Although intensive beam modeling would have been needed in any case, it was all the more important for WMAP because of the cooldown distortions. Due to composite CTE mismatches, the in-flight variations across the center of the primary were approximately 0.5–1 mm, as shown in Hill et al. (2009). To understand the distortions, we developed a model in which the surfaces were parametrized by over 400 Fourier modes (Jarosik et al. 2007; Hill et al. 2009). For each set of parameters, the full physical optics solution weighted by the measured passband was computed for all feeds. We then compared the physical optics prediction to measurements of Jupiter. The solution required running for many months on a 100 processors on Silicon Graphics Origin 300 machines. The reduced χ^2 of the fit are typically < 1.1 , suggesting quite a good fit given the overall complexity of the system coupled with multiple precise measurements of Jupiter. With the combination of the measurements and the model, the beam could be characterized at the -40 to -50 dB level, and the beam solid angles were determined to better than 1% (Hill et al. 2009).

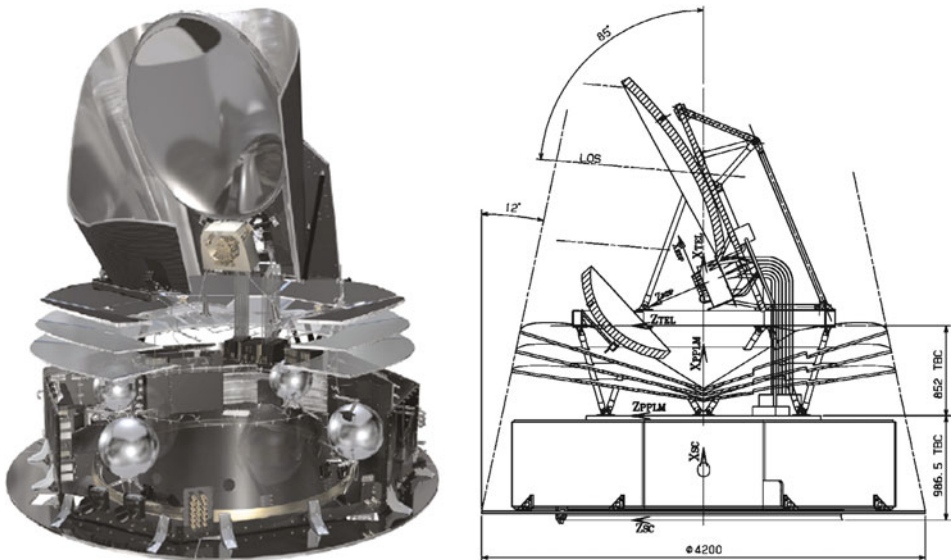
5.4 Planck

Planck is significantly more complex and sensitive than WMAP. Thus, although it was conceived at roughly the same time, it took longer to build. Planck combines two different technologies in one focal plane. The low-frequency instrument (LFI) uses coherent detectors cooled to 20 K operating between 30 and 70 GHz. The high-frequency instrument uses bolometric detectors cooled to 0.1 K operating between 90 and 900 GHz. An attractive aspect of the design is that it covers a very wide frequency range in one spacecraft. Much of the telescope optimization that was done for WMAP, and discussed earlier, was independently carried out for Planck. However, because of Planck’s enormous frequency range and much higher sensitivity ( Table 10-5), many of the optical system specifications were considerably more demanding than similar ones for WMAP.

Planck is the first CMB satellite to use a single telescope. This choice was driven by the bolometers in the HFI, the most sensitive detectors aboard Planck. The LFI instrument uses a differential receiver except that one of the inputs is terminated in a cold load as opposed to the sky as with WMAP. For bolometers, the analogue to a differential microwave receiver is a Fourier transform spectrometer (FTS). Not only would an FTS be more cumbersome and costly to build, but the intrinsically large FTS bandwidth is not amenable to single-mode optics or to minimizing the photon background. A single telescope was the natural choice.

The Planck telescope, shown in [Fig. 10-15](#), is an aplanatic Gregorian (Tauber et al. 2010a) and is roughly 20% larger than WMAP's. Planck's primary is 1.56×1.89 m and the secondary is 1.05×1.10 m. Thus, at a given frequency, it has about 20% more resolution. As with WMAP, the drivers for the design are to get the maximum number of feeds in the focal plane, with the most symmetric beams, in the most compact telescope. Planck had the additional challenge of supporting two different technologies in the focal plane operating at different temperatures. The in-flight surface accuracy is significantly better than WMAP's.

Planck HFI observations were completed in early 2012 when the instrument ran out of liquid cryogenics as expected. Early publications from Planck have demonstrated the excellent performance of the instrument, and more results and CMB maps will be released by the Planck team in the coming years.



■ Fig. 10-15

A cutaway view and line drawing of Planck. The focal plane is located just below the primary reflector. Planck spins about a vertical axis. As with WMAP, Planck is located at the second Lagrange point, roughly 1.5×10^6 km away from Earth. In this orientation, the Sun, Earth, and Moon are in the direction of the *bottom* of the page. Note the open ground shield around the optical system and the relatively large secondary reflector (Figures from Tauber et al. (2010a, b))

5.5 A Future Satellite

Because a satellite is the ideal platform for measuring the CMB, considerable effort has gone into designing the next generation instrument. After Planck, there will be little motivation for measuring the primary temperature anisotropy for $\ell \lesssim 2,500$ if the fluctuations turn out to be Gaussian to the limits of Planck's detector noise. However, there is quite a bit more to be learned from measuring the polarization, particularly the B-mode, as described in [Sect. 1.5](#). At the time of this writing, there are two relatively mature US satellite concepts: EPIC (Bock et al. 2008, 2009) and PIXIE (Kogut et al. 2011). These are very different missions. EPIC is based on the next generation single-moded detectors operating in the limit of CMB photon noise. In the EPIC-IM design, about 11,000 transition edge sensor bolometers are fed by a 4-K-cooled crossed-Dragone telescope that provides sensitivity to $\ell \gtrsim 1,500$. The detectors are cooled to ~ 100 mK by means of a continuously operating adiabatic demagnetization refrigerator. PIXIE, on the other hand, operates in the limit of many modes that are measured by just four detectors at the output of a polarizing fourier transform Spectrometer (FTS). Its sensitivity to B-mode science is at $\ell \lesssim 200$. The frequency spectrum of the anisotropy is determined by scanning the reflectors of the FTS much in the same ways as was done by COBE/FIRAS in its absolute measurement on the sky and by BAM (Tucker et al. 1997) in its search for the spectrum of the anisotropy.

A European collaboration has proposed the CORE mission concept (The CORE Collaboration et al. 2011). It is based on ~ 1.5 -m-diameter rotating half-wave plate as the first optical element that feeds a two-mirror system. The primary reflector is ~ 1.8 m diameter. The resolution varies between 23 and 1.3 arc-minutes for 45 and 795 GHz, respectively. Detection is based on feedhorn-coupled superconducting detectors to maintain control of systematics and achieve single-mode coupling with high sensitivity.

Acknowledgments

The authors are collaborators on just a subset of the instruments discussed in this chapter. We have learned about other optical system through papers and talking with colleagues, though all errors are of course ours. We would especially like to thank Elia Battistelli, Cynthia Chiang, Nils Halverson, Bill Jones, Akito Kusaka, Jeff McMahon, Lucio Piccirillo, Jon Sievers, Suzanne Staggs, Ed Wollack, and Sasha Zhiboedov for discussions and suggestions that improved this chapter. Ed Wollack in particular made numerous helpful comments. We also thank Chaoyun Bao, Angela Glenn, Michael Milligan, and Keith Thompson for help with the figures.

References

- Ade, P., et al. 2008, *ApJ*, 674, 22
 Ade, P. A. R., et al. 2010, *A&A*, 520, A11
 Aikin, R. W., et al. 2010, in *SPIE Conf. Ser.* 7741
 Ali-Haïmoud, Y., Hirata, C. M., & Dickinson, C. 2009, *MNRAS*, 395, 1055
 Alsop, D. C., et al. 1992, *ApJ*, 395, 317
 Arnold, K., et al. 2010, in *SPIE Conf. Ser.* 7741
 Baker, J. C., et al. 1999, *MNRAS*, 308, 1173
 Balanis, D. 2005, *Antenna Theory: Analysis and Design* (3rd ed.; Hoboken, NJ: Wiley-Interscience)
 Barkats, D., et al. 2005, *ApJL*, 619, L127
 Barnes, C., et al. 2002, *ApJS*, 143, 567
 Basak, S., Hajian, A., & Souradeep, T. 2006, *Phys. Rev. D*, 74, 021301

- Bennett, C. L. et al. 2003, *ApJ*, 583, 1
- Benôt, A., et al. 2003, *A&A*, 399, L19
- Bock, J., et al. 2008, *ArXiv e-prints*
- Bock, J., et al. 2009, *ArXiv e-prints*
- Boggess, N. W., et al. 1992, *ApJ*, 397, 420
- Bond, J. R., & Efstathiou, G. 1984, *ApJL*, 285, L45
- Born, M., & Wolf, E. 1980, *Principles of Optics* (6th ed.; Oxford/New York: Pergamon Press)
- Boughn, S. P., Fram, D. M., & Patridge, R. B. 1971, *ApJ*, 165, 439
- Boughn, S. P., Cheng, E. S., Cottingham, D. A., & Fixsen, D. J. 1992, *ApJL*, 391, L49
- Brown, K. W., & Prata, A. 1994, *IEEE Trans. Antennas Propag.*, 42, 1145
- Caderni, N., Fabbri, R., de Cosmo, V., Melchiorri, B., Melchiorri, F., & Natale, V. 1977, *Phys. Rev. D*, 16, 2424
- Carlstrom, J. E., et al. 2011, *PASP*, 123, 568
- Carpenter, R. L., Gulkis, S., & Sato, T. 1973, *ApJL*, 182, L61
- Chang, S., & Prata, A. 1999, *Antennas Propag. Soc. Int. Symp. IEEE*, 2, 1140
- Chang, S., & Prata, A. 2005, *J. Opt. Soc. Am. A*, 22, 2454
- Charlassier, R. 2008, *ArXiv e-prints*
- Cheng, E. S., Saulson, P. R., Wilkinson, D. T., & Corey, B. E. 1979, *ApJL*, 232, L139
- Cheng, E. S., et al. 1994, *ApJL*, 422, L37
- Chiang, H. C., et al. 2010, *ApJ*, 711, 1123
- Church, S. E. 1995, *MNRAS*, 272, 551
- Clarricoats, P., & Olver, A. 1984, *IEE Electromagn. Waves Ser., Corrugated Horns for Microwave Antennas* (London: P. Peregrinus on behalf of the Institution of Electrical Engineers)
- Conklin, E. K. 1969a, Ph.D. thesis, Stanford University
- Conklin, E. K. 1969b, *Nature*, 222, 971
- Conklin, E. K., & Bracewell, R. N. 1967, *Nature*, 216, 777
- Crill, B. P., et al. 2003, *ApJS*, 148, 527
- dall'Oglio, G., & de Bernardis, P. 1988, *ApJ*, 331, 547
- Das, S., et al. 2011, *ApJ*, 729, 62
- Davies, R. D., Lasenby, A. N., Watson, R. A., Daintree, E. J., Hopkins, J., Beckman, J., Sanchez Almeida, J., & Rebolo, R. 1987, *Nature*, 326, 462
- de Bernardis, P., et al. 1990, *ApJL*, 360, L31
- de Bernardis, P., et al. 2000, *Nature*, 404, 955
- de Oliveira-Costa, A., Kogut, A., Devlin, M. J., Netterfield, C. B., Page, L. A., & Wollack, E. J. 1997, *ApJL*, 482, L17
- de Oliveira-Costa, A., Devlin, M. J., Herbig, T., Miller, A. D., Netterfield, C. B., Page, L. A., & Tegmark, M. 1998, *ApJL*, 509, L77
- Dicker, S. R., et al. 1999, *MNRAS*, 309, 740
- Dragone, C. 1978, *AT T Tech. J.*, 57, 2663
- Dragone, C. 1982, *IEEE Trans. Antennas Propag.*, 30, 331
- Dragone, C. 1983a, *IEEE Trans. Antennas Propag.*, 31, 764
- Dragone, C. 1983b, *Electron. Lett.*, 19, 1061
- Dragovan, M., Ruhl, J. E., Novak, G., Platt, S. R., Crone, B., Pernic, R., & Peterson, J. B. 1994, *ApJL*, 427, L67
- Draine, B. T., & Lazarian, A. 1998, *ApJ*, 508, 157
- Draine, B. T., & Lazarian, A. 1999, *ApJ*, 512, 740
- Dudok, E. W. M & Fasold, E. 1986, *Analysis of Compact Antenna Test Range Configurations, JINA' 86 - International Symposium on Antennas*
- Epstein, E. E. 1967, *ApJL*, 148, L157
- Essinger-Hileman, T. 2011, Ph.D. thesis, Princeton University
- Fabbri, R., Guidi, I., Melchiorri, F., & Natale, V. 1980a, *Phys. Rev. Lett.*, 44, 1563
- Fabbri, R., Melchiorri, B., Melchiorri, F., Natale, V., Caderni, N., & Shivanandan, K. 1980b, *Phys. Rev. D*, 21, 2095
- Farese, P. C., et al. 2003, *New Astron. Rev.*, 47, 1033
- Filippini, J. P., et al. 2010, in *SPIE Conf. Ser.* 7741
- Finlay, C. C., et al. 2010, *Geophys. J. Int.*, 183, 1216
- Fixsen, D. J., & Mather, J. C. 2002, *ApJ*, 581, 817
- Fixsen, D. J., Cheng, E. S., & Wilkinson, D. T. 1983, *Phys. Rev. Lett.*, 50, 620
- Fowler, J. W., et al. 2005, *ApJS*, 156, 1
- Fowler, J. W., et al. 2007, *Appl. Opt.*, 46, 3444
- Fowler, J. W., et al. 2010, *ApJ*, 722, 1148
- Gaier, T., Schuster, J., Gundersen, J., Koch, T., Seiffert, M., Meinhold, P., & Lubin, P. 1992, *ApJL*, 398, L1
- Gaier, T., Lawrence, C. R., Seiffert, M. D., Wells, M. M., Kangaslahti, P., & Dawson, D. 2003, *New Astron. Rev.*, 47, 1167
- Ganga, K., Ratra, B., Church, S. E., Sugiyama, N., Ade, P. A. R., Holzappel, W. L., Mauskopf, P. D., & Lange, A. E. 1997, *ApJ*, 484, 517
- Graham, R. 1973, in *IEEE International Conference on Radar - Present and Future (IEEE)*, 134-139
- Grainge, K., et al. 2003, *MNRAS*, 341, L23
- Griffin, M. J., Bock, J. J., & Gear, W. K. 2002, *Appl. Opt.*, 41, 6543
- Halpern, M., Benford, R., Meyer, S., Muehler, D., & Weiss, R. 1988, *ApJ*, 332, 596
- Halverson, N. W., et al. 2002, *ApJ*, 568, 38
- Hanany, S., & Marrone, D. P. 2002, *Appl. Opt.*, 41, 4666
- Hanany, S., & Rosenkranz, P. 2003, *New Astron. Rev.*, 47, 1159
- Hanany, S., et al. 2000, *ApJL*, 545, L5
- Hanson, D., Lewis, A., & Challinor, A. 2010, *Phys. Rev. D*, 81, 103003
- Hecht, E. 1987, *Optics* (Reading, MA: Addison-Wesley)

- Hedman, M. M., Barkats, D., Gundersen, J. O., Staggs, S. T., & Winstein, B. 2001, *ApJL*, 548, L111
- Henry, P. S. 1971, *Nature*, 231, 516
- Hill, R. S., et al. 2009, *ApJS*, 180, 246
- Hinderks, J. R., et al. 2009, *ApJ*, 692, 1221
- Hinshaw, G., et al. 2007, *ApJS*, 170, 288
- Hu, W., Hedman, M. M., & Zaldarriaga, M. 2003, *Phys. Rev. D*, 67, 043004
- Imbriale, W. A., Gundersen, J., & Thompson, K. L. 2011, *IEEE Tran. Antennas Propag.*, 59, 1972
- Irwin, K., & Hilton, G. 2005, *Transition-edge sensors, in Cryogenic Particle Detection* (Berlin/New York: Springer)
- Jarosik, N., et al. 2003, *ApJS*, 145, 413
- Jarosik, N., et al. 2007, *ApJS*, 170, 263
- Johnson, B. R., et al. 2007, *ApJ*, 665, 42
- Jones, W. C. 2005, Ph.D. thesis, California Institute of Technology
- Kamionkowski, M., Kosowsky, A., & Stebbins, A. 1997, *Phys. Rev. D*, 55, 7368
- Keating, B. G., Timbie, P. T., Polnarev, A., & Steinberger, J. 1998, *ApJ*, 495, 580
- Keating, B. G., O'Dell, C. W., de Oliveira-Costa, A., Klawikowski, S., Stebor, N., Piccirillo, L., Tegmark, M., & Timbie, P. T. 2001, *ApJL*, 560, L1
- Keisler, R., et al. 2011, *ApJ*, 743, 28
- Klypin, A. A., Strukov, I. A., & Skulachev, D. P. 1992, *MNRAS*, 258, 71
- Knoke, J. E., Partridge, R. B., Ratner, M. I., & Shapiro, I. I. 1984, *ApJ*, 284, 479
- Kogut, A., et al. 1992, *ApJ*, 401, 1
- Kogut, A., Banday, A. J., Bennett, C. L., Górski, K. M., Hinshaw, G., & Reach, W. T. 1996, *ApJ*, 460, 1
- Kogut, A., et al. 2011, *J. Cos. & Par. Ast.*, 7, 25
- Komatsu, E., et al. 2011, *ApJS*, 192, 18
- Korsch, D. 1991, *Reflective Optics* (Boston: Academic)
- Kovac, J. M., Leitch, E. M., Pryke, C., Carlstrom, J. E., Halverson, N. W., & Holzzapfel, W. L. 2002, *Nature*, 420, 772
- Kuo, C. L., et al. 2004, *ApJ*, 600, 32
- Kuo, C. L., et al. 2008, in *SPIE Conf. Ser. 7020*
- Lamarre, J. M. 1986, *Appl. Opt.*, 25, 870
- Lasenby, A. N., & Davies, R. D. 1983, *MNRAS*, 203, 1137
- Lay, O. P., & Halverson, N. W. 2000, *ApJ*, 543, 787
- Ledden, J. E., Broderick, J. J., Brown, R. L., & Condon, J. J. 1980, *AJ*, 85, 780
- Leitch, E. M., Readhead, A. C. S., Pearson, T. J., & Myers, S. T. 1997, *ApJL*, 486, L23
- Leitch, E. M., Readhead, A. C. S., Pearson, T. J., Myers, S. T., Gulkis, S., & Lawrence, C. R. 2000, *ApJ*, 532, 37
- Levy, A. R., et al. 2008, *ApJS*, 177, 419
- Love, A., *Antennas, I., & Society*, P. 1978, *IEEE Press Sel. Repr. Ser., Reflector Antennas* (New York: IEEE Press)
- Lubin, P. M., Epstein, G. L., & Smoot, G. F. 1983, *Phys. Rev. Lett.*, 50, 616
- Maffei, B., et al. 2010, *A&A*, 520, A12
- Mandolesi, N., Calzolari, P., Cortiglioni, S., Delpino, F., & Sironi, G. 1986, *Nature*, 319, 751
- Mather, J. C. 1982, *Appl. Opt.*, 21, 1125
- Mather, J., Hinshaw, G., & Page, L. 2012, *Cosmic Microwave Background. Planets, Stars and Stellar Systems. Vol. 6*, Springer
- McMahon, J. J., et al. 2009, in *AIP Conf. Ser. 1185*, ed. B. Young, B. Cabrera, & A. Miller, 511–514
- McMahon, J., et al. 2012, *ArXiv e-prints*
- Meinhold, P., & Lubin, P. 1991, *ApJL*, 370, L11
- Meinhold, P. R., Chinguanco, A. O., Gundersen, J. O., Schuster, J. A., Seiffert, M. D., Lubin, P. M., Morris, D., & Vilella, T. 1993, *ApJ*, 406, 12
- Meinhold, P. R., et al. 2005, *ApJS*, 158, 101
- Mennella, A., et al. 2011, *ArXiv e-prints*
- Meyer, S. S., Cheng, E. S., & Page, L. A. 1991, *ApJL*, 371, L7
- Miller, A. D., et al. 1999, *ApJL*, 524, L1
- Mizuguchi, Y., & Yokoi, H. 1974, *International Conv. of IECE, Japan*, 801
- Mizuguchi, Y., Akagawa, M., & Yokoi, H. 1978, *Electron. Commun. Jpn.*, 61, 58
- Mizugutch, Y., & Yokoi, H. 1975, *Trans of IECE of Japan*, 58–3
- Mizugutch, Y., Akagawa, M., & Yokoi, H. 1976, in *Digest of 1976 AP-S International Symposium on Antennas and Propagation* (New York: IEEE)
- Muehlner, D. 1977, in *Astrophys. Space Sci. Libr.* 63, *Infrared and Submillimeter Astronomy*, ed. G. G. Fazio (Dordrecht/Boston: D. Reidel Pub. Co.), 143–152
- Nanos, G. P. Jr., 1979, *ApJ*, 232, 341
- Niemack, M. D., et al. 2008, *J. Low Temp. Phys.*, 151, 690
- Niemack, M. D., et al. 2010, in *SPIE Conf. Ser.* 7741
- O'Brient, R., et al. 2008, *J. Low Temp. Phys.*, 151, 459
- O'Dea, D., Challinor, A., & Johnson, B. R. 2007, *MNRAS*, 376, 1767
- Olver, A. D. 1991, in *Antennas and Propagation, 1991. ICAP 91., Seventh International Conference on (IEE)*, Vol. 1 (London: IEE), 99–108
- Olver, A., Clarricoats, P., Kishk, A., & Shafai, L. 1994, *Microwave Horns and Feeds* (London: IEEE Press and IEE)
- O'Sullivan, C., et al. 1995, *MNRAS*, 274, 861
- O'Sullivan, C., et al. 2008, *Infrared Phys. Technol.*, 51, 277
- Padin, S., et al. 2001, *ApJL*, 549, L1
- Padin, S., et al. 2008, *Appl. Opt.*, 47, 4418

- Page, L. A., Cheng, E. S., & Meyer, S. S. 1990, *ApJL*, 355, L1
- Page, L. A., Cheng, E. S., Golubovic, B., Meyer, S. S., & Gundersen, J. 1994, *App. Opt.*, 33, 11
- Page, L., et al. 2003, *ApJ*, 585, 566
- Pardo, J. R., Cernicharo, J., & Serabyn, E. 2001, *IEEE Trans. Antennas Propag.*, 49, 1683
- Pariiskii, Y. N., & Pyatunina, T. B. 1971, *Sov. Astro.*, 14, 1067
- Parijskij, Y. N. 1973, *ApJL*, 180, L47
- Partridge, R. B. 1980, *ApJ*, 235, 681
- Partridge, R. B. 1995, *3K: The Cosmic Microwave Background Radiation* (Cambridge: Cambridge University Press)
- Pascale, E., et al. 2008, *ApJ*, 681, 400
- Pearson, T. J., et al. 2003, *ApJ*, 591, 556
- Peebles, P. J. E. 1994, *ApJL*, 432, L1
- Peebles, P. J. E., Page, L., & Partridge, B. 2009, *Finding the Big Bang* (Cambridge: Cambridge University Press)
- Peterson, J. B., et al. 2000, *ApJL*, 532, L83
- Piccirillo, L. 1991, *Rev. Sci. Instrum.*, 62, 1293
- Piccirillo, L., & Calisse, P. 1993, *ApJ*, 411, 529
- Piccirillo, L., et al. 1997, *ApJL*, 475, L77
- Piccirillo, L., et al. 2008, in *SPIE Conf. Ser. 7020*
- Planck HFI Core Team, et al. 2011, *ArXiv e-prints*
- Pospieszalski, M. W. 1992, in *IEEE MTT-S Digest*, 1369
- QUIET Collaboration, et al. 2011, *ApJ*, 741, 111
- Rabii, B., et al. 2006, *Rev. Sci. Instrum.*, 77, 071101
- Rahmat-Samii, Y., Imbriale, W., & Galindo-Isreal, V. 1995, *DADRA, YRS Associates*, rahmat@ee.ucla.edu
- Readhead, A. C. S., Lawrence, C. R., Myers, S. T., Sargent, W. L. W., Hardebeck, H. E., & Moffet, A. T. 1989, *ApJ*, 346, 566
- Reichborn-Kjennerud, B., et al. 2010, in *SPIE Conf. Ser. 7741*
- Richards, P. L. 1994, *J. Appl. Phys.*, 76, 1
- Rudnick, L. 1978, *ApJ*, 223, 37
- Schaffer, K. K., et al. 2011, *ApJ*, 743, 90
- Schlaerth, J. A., et al. 2010, in *SPIE Conf. Ser. 7741*
- Scott, P. F., et al. 1996, *ApJL*, 461, L1
- Seielstad, G. A., Masson, C. R., & Berge, G. L. 1981, *ApJ*, 244, 717
- Seo, E. S., et al. 2008, *Adv. Space Res.*, 42, 1656
- Sheehy, C. D., et al. 2010, in *SPIE Conf. Ser. 7741*
- Shimon, M., Keating, B., Ponthieu, N., & Hivon, E. 2008, *Phys. Rev. D*, 77, 083003
- Shirokoff, E., et al. 2011, *ApJ*, 736, 61
- Singal, J., et al. 2011, *ApJ*, 730, 138
- Smoot, G. F., Gorenstein, M. V., & Muller, R. A. 1977, *Phys. Rev. Lett.*, 39, 898
- Smoot, G., et al. 1990, *ApJ*, 360, 685
- Smoot, G. F., et al. 1991, *ApJL*, 371, L1
- Staniszewski, Z., et al. 2009, *ApJ*, 701, 32
- Stankevich, K. S. 1974, *Sov. Astro.*, 18, 126
- Staren, J., et al. 2000, *ApJ*, 539, 52
- Strukov, I. A., & Skulachev, D. P. 1984, *Sov. Astron. Lett.*, 10, 1
- Strukov, I. A., & Skulachev, D. P. 1986, *Itogi Nauki i Tekhniki Seriya Astronomiia*, 31, 37
- Strukov, I. A., Skulachev, D. P., Boyarskiy, M. N., & Tkachev, A. N. 1987, *JPRS Rep. Sci. Technol. USSR Space*, 3, 59
- Strukov, I. A., Skulachev, D. P., & Klypin, A. A. 1988, in *IAU Symp. 130, Large Scale Structures of the Universe*, ed. J. Audouze, M.-C. Pelletan, A. Szalay, Y. B. Zel'Dovich, & P. J. E. Peebles (Dordrecht/Boston: Kluwer), 27-+
- Su, M., Yadav, A. P. S., Shimon, M., & Keating, B. G. 2011, *Phys. Rev. D*, 83, 103007
- Sunyaev, R. A., & Zeldovich, I. B. 1980, *ARAA*, 18, 537
- Swetz, D. S., et al. 2011, *ApJS*, 194, 41
- Takahashi, Y. D., et al. 2010, *ApJ*, 711, 1141
- Tatarskii, V. I. 1961, *Wave Propagation in a Turbulent Medium* (New York: McGraw-Hill)
- Tauber, J. A., et al. 2010a, *A&A*, 520, A2+
- Tauber, J. A., et al. 2010b, *A&A*, 520, A1+
- Tegmark, M., & Zaldarriaga, M. 2009, *Phys. Rev. D*, 79, 083530
- Tegmark, M., & Zaldarriaga, M. 2010, *Phys. Rev. D*, 82, 103501
- The CoRE Collaboration, et al. 2011, *ArXiv e-prints*
- The Qubic Collaboration, et al. 2011, *Astropart. Phys.*, 34, 705
- Timbie, P. T., & Wilkinson, D. T. 1988, *Rev. Sci. Instrum.*, 59, 914
- Timbie, P. T., & Wilkinson, D. T. 1990, *ApJ*, 353, 140
- Timbie, P. T., et al. 2006, *New Astro. Rev.*, 50, 999
- Toral, M. A., Ratliff, R. B., Lecha, M. C., Maruschak, J. G., & Bennett, C. L. 1989, *IEEE Trans. Antennas Propag.*, 37, 171
- Tran, H., & Page, L. 2009, *J. Phys. Conf. Ser.*, 155, 012007
- Tran, H., Lee, A., Hanany, S., Milligan, M., & Renbarger, T. 2008, *App. Opt.*, 47, 103
- Tran, H., et al. 2010, in *SPIE Conf. Ser. 7731*
- Tucker, G. S., Griffin, G. S., Nguyen, H. T., & Peterson, J. B. 1993, *ApJL*, 419, L45+
- Tucker, G. S., Gush, H. P., Halpern, M., Shinkoda, I., & Towlson, W. 1997, *ApJL*, 475, L73
- Uson, J. M., & Wilkinson, D. T. 1982, *Phys. Rev. Lett.*, 49, 1463
- Vokurka, V. J. 1980, *Compact Antenna Range Performance at 70 GHz*, *IEEE Int. Symp. Ant. and Prop.*, 260-263
- Watson, R. A., Gutierrez de La Cruz, C. M., Davies, R. D., Lasenby, A. N., Rebolo, R., Beckman, J. E., & Hancock, S. 1992, *Nature*, 357, 660

- Weiss, R. 1980, *ARAA*, 18, 489
- Welton, W., & Winston, R. 1978, *The Optics of Non-imaging Concentrators* (New York: Academic)
- Wilkinson, D. T., & Partridge, R. B. 1967, *Nature*, 215, 719
- Wilson, R. W., & Penzias, A. A. 1967, *Science*, 156, 1100
- Wollack, E. J., Jarosik, N. C., Netterfield, C. B., Page, L. A., & Wilkinson, D. 1993, *ApJL*, 419, L49
- Yadav, A. P. S., Su, M., & Zaldarriaga, M. 2010, *Phys. Rev. D*, 81, 063512
- Yoon, K. W., et al. 2006, in *SPIE Conf. Ser.* 6275
- Yoon, K. W., et al. 2009, in *AIP Conf. Ser.* 1185, ed. B. Young, B. Cabrera, & A. Miller, 515–518
- Zaldarriaga, M., & Seljak, U. 1997, *Phys. Rev. D*, 55, 1830
- Zmuidzinas, J. 2003, *Appl. Opt.*, 42, 4989

11 Very-High-Energy Gamma-Ray Telescopes

Elizabeth Hays

NASA Goddard Space Flight Center, Greenbelt, MD, USA

1	<i>Introduction</i>	482
2	<i>Gamma Rays in Matter</i>	483
3	<i>Space-Based Gamma-Ray Telescopes</i>	485
4	<i>Ground-Based Gamma-Ray Telescopes</i>	486
4.1	Extensive Air Showers	487
4.2	Water Cherenkov Telescopes	490
4.3	WCTs: A Little History	494
4.4	Imaging Atmospheric Cherenkov Telescopes	494
4.5	IACTs: A Little History	504
	<i>References</i>	504

Abstract: Telescopes capable of collecting the highest energy light, called gamma radiation, explore a region of the electromagnetic spectrum that was largely inaccessible until recently. Observations at the highest energies reveal the existence and properties of extreme sites in the Universe. In this chapter, I cover the techniques used to detect photons at frequencies above 10^{23} Hz, photon energy >100 MeV, in the so-called pair production regime. Telescopes employing particle and optical detection methods are used in space or the upper atmosphere as well as on the ground. Gamma-ray astrophysics is a young field with only one or two generations of instruments completed for each of several techniques. In the past decade, telescopes throughout this band have generated significantly deeper and more complete surveys and catalogs of the high- and very-high-energy sky. Already the results are spectacular, but many challenges remain for enhancing the performance of high- and very-high-energy gamma-ray telescopes.

Keywords: Cherenkov radiation, Cherenkov telescope, Extensive air shower, Gamma ray, Imaging atmospheric, Pair production, Water Cherenkov telescope

List of Abbreviations: *eV*, electronvolt; *GRB*, Gamma-ray burst; *HAWC*, High altitude water Cherenkov; *HE*, High energy; *H.E.S.S.*, High energy stereoscopic system; *IACT*, Imaging atmospheric Cherenkov telescope; *LAT*, Large area telescope; *PMT*, Photomultiplier tube; *VERITAS*, Very energetic radiation imaging telescope array system; *VHE*, Very high energy; *WCT*, Water Cherenkov telescope

1 Introduction

The highest energy photons are known as gamma radiation. This refers to an enormous band of the electromagnetic spectrum that spans many orders of magnitude in frequency beginning around 10^{18} Hz or about 10^5 electronvolts (eV) in the energy units commonly used in this regime. High-energy (HE; ~ 100 MeV– 100 GeV) and very-high-energy (VHE; ~ 100 GeV– 100 TeV) photons provide a tantalizing but elusive view of extremely powerful processes at work in the Universe. Extended jets from black holes at the center of galaxies, expanding shockwaves from supernova explosions, and magnetospheres of rapidly rotating neutron stars – these are a few of the environments that supply conditions that can accelerate matter to relativistic speeds. Gamma rays are an inevitable by product and therefore a telltale signature of particle acceleration at these sites. The sites of gamma-ray emission indicate astrophysical accelerators that exceed the capabilities of laboratories on Earth. Astrophysical gamma-ray signals have been observed above 10 TeV and even approaching 100 TeV ($\sim 10^{29}$ Hz) from supernova remnants in our Galaxy (Aharonian et al. 2007). These photons are among the most energetic recorded from distinct objects, and they indicate the presence of particles accelerated to similar energy. Recent observations of intense flares from the Crab Nebula indicate that somewhere within the zone where the wind from the powerful central pulsar meets the surrounding medium, electrons can attain energies in excess of 1 PeV (Abdo et al. 2011).

Gamma rays are used to map the distribution of the ubiquitous accelerated electrons and nuclei, the cosmic rays, that permeate galaxies. Cosmic-ray interactions with the interstellar medium create a gamma-ray glow that measures the cosmic-ray content of galaxies based on their gamma-ray luminosity (Abdo et al. 2010). Our own Milky Way Galaxy is the brightest feature in the high-energy gamma-ray sky. Additionally, our Galaxy has huge lobes of gamma-ray emission extending away from the central black hole. This feature, likely to be a “bubble” of energetic electrons, has been observed for the first time using gamma rays and is

still being understood (Su et al. 2010). Features such as these are best observed by very wide-field instruments and require substantial reduction of foregrounds and backgrounds.

Although there are many steady sources of gamma-ray emission, some of the most exciting observations concern transient phenomena. Gamma-ray bursts (GRBs), extremely fast and bright flashes linked to supernovae and mergers of compact objects, are inherently gamma-ray phenomena. The gamma-ray observations are typically the first indications used to locate these events and point other telescopes toward them. In GRBs, the gamma rays are the defining feature of an event that may last only a few seconds. The high-energy emission provides a critical test for the engines that drive the explosions and their expansion into the surrounding environment. GRBs are among the most distant objects observed, some at more than 13 billion light-years distant. Such succinct events at high energy and large distances provide an extremely sensitive test of our understanding of space and time (Abdo et al. 2009). The short-lived nature of these events makes wide-field observations and rapid slewing of pointed instruments a necessity.

Additionally, since very-high-energy gamma rays are observed in an energy regime that has not been explored in laboratories, they are an ideal tool to search for unexplained phenomena that could indicate new and exotic physics. As ground-based accelerator facilities search to confirm and extend the standard model and possibly discover the characteristics of dark matter, potential gamma-ray signatures remain a promising avenue for studying dark matter in astrophysical settings. In particular, current searches include seeking the unexplained presence of gamma rays from regions with high dark matter content or unexplained gamma-ray spectral features at a potential dark matter particle mass scale. These types of searches require careful understanding of the telescope performance and the gamma-ray sky.

The varied characteristics of gamma-ray sources demand a corresponding variety in observing platforms to fully exploit the astrophysical phenomena accessible throughout the waveband. A suite of telescopes with a combination of high sensitivity, good resolution, superb timing, broad energy coverage, and both broad and frequent sky coverage as well as fast response are key to probing the contents of the gamma-ray sky and the characteristics of its denizens.

Observing gamma rays requires unique telescopes and innovative techniques. Gamma radiation has very short wavelengths, at the scale of atomic nuclei and smaller, and falls in the regime where light is treated as a particle instead of a wave. The reflection and refraction techniques applicable at lower energies are not useful because instead of reflecting from a surface, a gamma ray passes through, scatters, or is destroyed within the material of a traditional reflector. However, the interaction products are not subtle and can be used to detect and characterize the gamma ray. Although several interaction processes occur throughout the large range of energy contained in the gamma-ray band, the telescopes described in this article specifically exploit the pair production interaction, the most common process for the highest energy gamma rays to undergo when encountering a material.

2 Gamma Rays in Matter

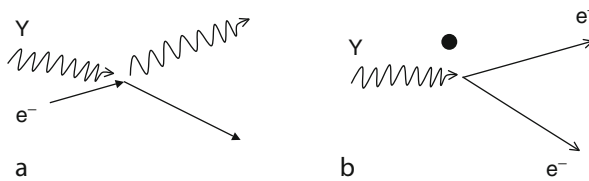
Before digging into the details of the telescopes, it is important to understand the properties of gamma rays.¹ They interact readily with matter. Within a few mm of lead or a few km of air, a gamma ray will be scattered, absorbed, or converted into particles. This has important

¹The Particle Data Group provides an excellent collection of references and reviews for many of the properties and measurements necessary to understand the interactions of gamma rays in detectors (Berlinger et al. 2012).

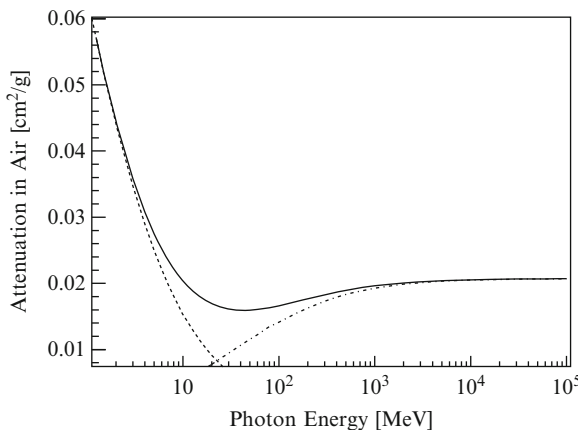
implications for gamma-ray telescopes. Because of these destructive interactions with matter, telescopes do not focus gamma rays, but instead look for signatures of their passage through the telescope.

Energetic photons interact electromagnetically with atoms through several processes. The dominant interaction mode depends on the photon energy and the properties of the material impacted. At the low energy end of the gamma-ray band (also referred to as a hard x-ray), a photon may be coherently scattered and can be completely absorbed by an atom ejecting an electron, the photoelectric effect. A more energetic gamma-ray photon is not absorbed and may free an electron from an atom, Compton scattering (👉 Fig. 11-1). This process becomes important from a few MeV to a few hundred MeV. The cross section, indicating the probability of the interaction, depends on the density of electrons in a material and decreases as approximately the inverse of the photon energy (see 👉 Fig. 11-2).

In the MeV range and above, photons passing near the nucleus of an atom can convert into a matter-antimatter pair, most commonly an electron and a positron (👉 Fig. 11-1). The threshold for electron-positron pair production is twice the rest mass energy of the electron, 1.022 MeV. The cross section for this process increases rapidly above the threshold energy and then remains




👉 Fig. 11-1 Interactions of high- and very-high-energy photons with matter. (a) Compton scattering. (b) Pair production



👉 Fig. 11-2 Attenuation coefficient for photons in air (Berger et al. 2010). The *solid line* gives the total attenuation, while the *dashed* and *dot-dashed lines* give the contributions from Compton scattering and pair production, respectively

roughly constant with increasing energy. This means that pair production begins to dominate where the rate of Compton scattering drops, typically around 100 MeV–1 GeV, depending on the medium. Pair production plays a more notable role below 1 GeV in materials with higher atomic number.

Photon interactions with matter are characterized by an attenuation coefficient, or cross section, that depends on the material properties (atomic number and mass) and the photon energy. The attenuation for air is shown in  *Fig. 11-2* to illustrate the energy dependence of several important interaction processes. The attenuation coefficient is given in inverse-length/density and indicates the depth at which a photon loses 1/e of its energy, or in the case of discrete interactions, the depth at which the probability that the photon has not interacted is 1/e. In the pair production regime, photon attenuation is commonly talked about in terms of the radiation length for a material, which is the characteristic distance for an electron in the material to radiate a photon, 7/9 of the mean free path. A radiation length is about 37 g/cm² in air or 6.8 g/cm² in tungsten (atomic number 74). The type and configuration of materials used in constructing a gamma-ray telescope provide a palette for tuning its characteristics: where the gamma interactions occur, what the dominant interaction mechanism is, and how the secondary products are collected and observed.

3 Space-Based Gamma-Ray Telescopes

Gamma rays do not penetrate the atmosphere of the Earth deeply enough to reach the ground. Only high-altitude balloons or space telescopes observe them directly. The instruments induce the gamma-ray interaction and then record the secondary products to measure the properties of the original gamma ray. Either of the high-energy interaction processes, Compton scattering or pair production, can be leveraged to detect a gamma ray, but telescopes are optimized to favor one process or the other. The energy dependence of the probable occurrence of the interactions determines an approximate energy window for each technique. From a few hundred keV to 30 MeV (depending on the material), gamma rays most commonly undergo Compton scattering, and so Compton telescopes are designed to look for a scattered electron. At higher energies, where the cross section for the Compton process drops, pair-production telescopes measure an electron and positron pair generated within the instrument.

The Compton Gamma-Ray Observatory (CGRO), which operated from 1991 to 2000, carried an imaging Compton telescope, COMPTEL (Schoenfelder et al. 1993), that operated at energies from 1 to 30 MeV. A liquid scintillator measured the direction and energy of the recoil electron from a Compton-scattered gamma ray. An array of sodium iodide crystals absorbed the remaining energy of the scattered gamma ray. An interesting feature of this technique is that the direction of the scattered gamma ray and the energy of the recoil electron conspire to constrain the direction of the primary gamma ray to lie on a circle. When many gamma rays are observed from the same source and a good knowledge of the instrument response is applied, the intersection of the circles indicates the direction of the gamma-ray source. COMPTEL achieved an angular resolution of about 1°.

A second instrument on CGRO, the Energetic Gamma Ray Experiment Telescope (EGRET), exploited the pair production interaction to explore energies from 20 MeV to 30 GeV (Thompson et al. 1993; Esposito et al. 1999). EGRET consisted of a wire spark chamber to record the electron and positron pair, scintillating detectors to trigger the spark chamber readout, and

a calorimeter to measure the deposited energy. The tracking of the electron and positron pair in the spark chamber allowed the direction measurement of the original gamma ray, but the gas degraded over time, limiting the life of the instrument. EGRET and COMPTEL, like most space-borne gamma-ray detectors, carried plastic scintillator detectors to serve as anticoincidence detectors, recording the passage of charged background particles into the instrument and allowing them to be separated from gamma rays.

More recent pair-production telescopes have succeeded EGRET. The Large Area Telescope (LAT) on the Fermi Gamma-ray Space Telescope and the AGILE satellite (Astrorivelatore Gamma a Immagini Leggero) (Tavani et al. 2009) are both currently operating in the high-energy regime. In particular, the LAT has exceeded the energy range available to EGRET by observing gamma rays from 20 MeV to >300 GeV (Atwood et al. 2009). Instead of a gas tracker, the LAT uses silicon strip detectors interleaved with tungsten to convert the incoming photon, trigger readout, and track the electron and positron pair. The LAT also includes a plastic scintillator anticoincidence system and an imaging calorimeter. The configuration of the LAT tracker enables a very large field of view, ~ 2 sr. Although at the lower end of the energy range, the limitation to angular resolution lies in scattering of the generated electron; at higher energies, the resolution is only limited by the pitch of the silicon detectors and at a fraction of a degree provides groundbreaking images of the high-energy sky. While EGRET operated as a pointed telescope, focusing on specific targets, the LAT, possessing a wider field of view and primarily operating in a survey mode, achieves complete coverage of the sky in about 3 h.

The lack of focusing optics for gamma-ray telescopes means that the angular resolution for these techniques remains far from diffraction limitations and enhanced sensitivity poses a significant difficulty. The collection power of pair production telescopes, particularly at higher energies, is closely linked to the telescope size and the amount of mass that it contains. Both of these properties are constrained for space platforms. However, a significant benefit of particle detection telescopes is their acceptance of gamma rays arriving from many directions. The fundamental limit for the field of view is the portion of sky that is not occulted by the Earth. Realistically, the field of view is often smaller because of directional biases in the structure of an instrument or in how it selects photons and rejects backgrounds. However, this still allows for very large fields of view, often one or two steradians of sky, comparable to the human eye. Observations made by space-based gamma-ray telescopes are also primarily unaffected by features imposed by the atmosphere, a source of major systemic errors for instruments on the ground.

For the rest of this chapter I will focus on ground-based instruments. For more information about the fundamentals and history of Compton telescopes, see von Ballmoos et al. (1989). For information about pair-production telescopes, see, for example, Thompson et al. (1993), Esposito et al. (1999), and Atwood et al. (2009).

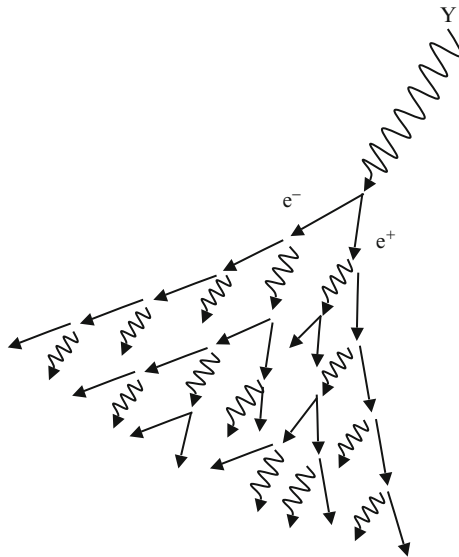
4 Ground-Based Gamma-Ray Telescopes

When gamma rays strike the Earth's atmosphere, they leave an unmistakable signature as their energy dissipates in an extended shower of electrons, positrons, and photons. The atmosphere plays a key role in all ground-based gamma-ray telescopes out of necessity. A TeV photon passes through a depth of about 48 g/cm^2 of air before undergoing pair production. The atmosphere

is $1,030 \text{ g/cm}^2$ thick. This means that gamma rays convert at altitudes high above sea level, typically above 10 km. The atmosphere both initiates the pair production interaction and transmits the by-products to instruments on the ground, making it the principal element of any ground-based gamma-ray telescope. In order to discuss telescopes situated on the ground, we must start with what happens above it.

4.1 Extensive Air Showers

Extensive air showers begin with a single photon. After the initial pair production reaction occurs, an electromagnetic cascade of electrons, positrons, and high-energy photons develops. The electron-positron pairs produced by TeV energy gamma rays are extremely energetic and do not proceed far from their creation point.² The electrons lose energy in the air via bremsstrahlung, literally “braking radiation,” and emit high-energy photons (☛ Fig. 11-1). A bremsstrahlung photon carries about half of the electron energy on average. This means that the emitted photons are still energetic enough to convert to electrons via pair production. As mentioned previously, the characteristic length scale for pair production and electron bremsstrahlung are similar, and so the reaction continues fairly equitably. The cascade develops multiplicatively with photons converting to pairs of electrons and positrons that then radiate to produce increasing numbers of particles and photons (☛ Fig. 11-3). The air shower peaks and dwindles when the electrons suffer enough loss through ionization that the radiated photon energies fall below the pair-production threshold.



☛ Fig. 11-3

Toy model of air shower interactions. The initial gamma ray converts into an electron and its antiparticle, the positron, which then generate bremsstrahlung gamma rays

²Further mention of electrons also refers to positrons unless indicated differently.

Further details of air shower development and modeling can be found in Gaisser (1991) or in the early treatment by Rossi and Greisen (1941). Here, I will consider some of the important features of air showers in the GeV-TeV domain that impact the design and performance of ground-based gamma-ray telescopes.

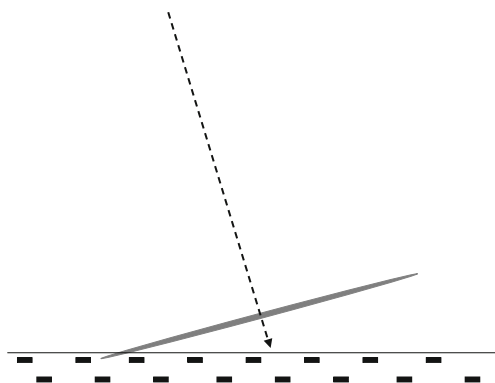
The pair production cross section in air is not strongly energy dependent above 1 GeV. The initial pair production interaction occurs at similar depths independently of the photon energy, on average at about 10 km above the ground. The development of the ensuing cascade, however, is strongly dependent on the energy of the original gamma ray. The more energy that must be dissipated, the more secondary particles are produced, and the more deeply the shower extends into the atmosphere. At energies exceeding 10 TeV, an appreciable number of energetic particles reach ground level, but at energies below 1 TeV, the particle density is much lower on average. As a consequence, instruments that measure the particle component of air showers benefit from being situated at high altitude.

The development of an air shower is a statistical process. Gamma rays of the same energy produce interactions over a range of altitudes and create air showers that penetrate to a range of depths. As a consequence, the longitudinal extent of the shower and the number of particles reaching an observatory on the ground fluctuate substantially from the average values. These observables offer an imprecise measure of the energy of the primary gamma ray. The lateral development of an air shower is also driven by statistical fluctuations in the cascade interactions that broaden the pool of secondary particles. Although the pair production and bremsstrahlung processes contribute to the angular spread, Compton scattering of the electrons dominates the transverse size. The transverse size of the shower sets the scale for the area on the ground that an instrument must cover to contain the shower.

The particle cascade continues for many iterations, often called generations, of the pair production and bremsstrahlung processes, but the shower front remains well defined. The thin disk of energetic secondary particles may cover several hundred meters at ground level, but remains only a few ns thick. In other words, a vertically propagating shower produces a signal that lasts for only a few ns. The sensors and electronic systems in gamma-ray telescopes must respond and sample data very quickly to distinguish an air shower from backgrounds that vary more slowly over time.

Information about the primary gamma ray is encoded in the air shower that it generates. The propagation direction of the shower remains largely true to the direction of the original gamma ray (ignoring secondary effects like refraction in the atmosphere and distortions from the Earth's magnetic field acting on the charged particles). The relatively narrow plane of the shower front at ground level provides a good indication of the origin of the gamma ray (► Fig. 11-4). To make a quick and highly idealized estimate of the angular resolution we might expect, consider having perfect knowledge of the shower front. A few ns for near-relativistic speed translates into about 1 m in width for the shower plane, which may extend over a diameter of about 200 m at 10 TeV. This gives an approximate angular error of 0.3° for a single photon.

The gamma-ray energy transfers into secondary particles and high-energy photons. The more energy that a gamma ray carries, the more particles that are generated, and the more deeply the shower penetrates into the atmosphere. An air shower can be characterized by the depth at which it has generated the maximum number of particles, the shower maximum. The doubling nature of the interactions makes the shower maximum proportional to the natural logarithm of the primary gamma-ray energy. Alternatively, a measurement of the extent and density of particles on the ground gives a much coarser indication of the energy. The statistical fluctuations for both the shower maximum and the number of particles reaching the



■ Fig. 11-4

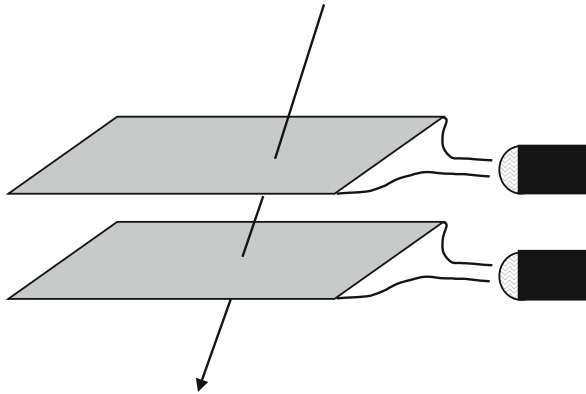
Cartoon illustrating the particle shower front, a thin pancake of surviving particles from the extended air shower, traveling along the direction of the original gamma ray at close to the speed of light and approaching an array of ground detectors

ground are large and can only indicate the gamma-ray energy in an average sense and not for single photons. In other words, the average number of particles for a large sample of gamma-ray air showers from a hypothetical source of 10 TeV photons differ from a source of 0.1 TeV photons. In reality, gamma ray sources emit over a continuum, and the business of distinguishing spectral shapes using the particle content of the air shower is far from simple.

One way to observe an air shower is to simply detect the relativistic particles that reach ground level. This is easy to do in a one-dimensional sense and can be accomplished by using plastic scintillator, which emits light when charged particles pass through it, coupled to fast light sensors and some basic electronics to count the flashes of light.³ Such a simple detector can measure the rate that air showers strike it. In combination with the area of sky that the detector observes and the efficiency for detecting showers passing through the scintillator, this rate can be converted into a flux of air showers. However, this simple counter does not provide the direction of the air showers on the sky or any estimate of their energy. To remedy this, an array of scintillator devices can be combined to provide knowledge of the shower extent and relative arrival times of the shower front on the ground (► Fig. 11-5).

There are significant challenges for doing gamma-ray astronomy with a simple scintillator array. Most importantly, the array must be large enough to collect an interesting number of gamma rays; a typical TeV shower covers several hundred meters on the ground. Also, some minimum number of particles must strike the scintillator array to allow a useful measurement of an air shower. That translates into a minimum energy for detectable gamma rays. Unless dense coverage is possible over a large area, this minimum energy is uncomfortably high compared to the bulk of expected emission from astrophysical sources. Additionally, we would really like to know the rate of gamma rays impacting the top of the atmosphere not just the rate of air showers reaching the ground. A gamma ray can strike the top of the atmosphere and generate a shower that does not reach the detector, creating a substantial inefficiency for this method, particularly for lower energy gamma rays. Detailed simulations of the atmosphere and the detector

³This device commonly uses pairs of scintillating plastic “paddles” and is often referred to as a muon telescope because it is most likely registering the penetrating muons generated in cosmic-ray-induced air showers.



■ Fig. 11-5

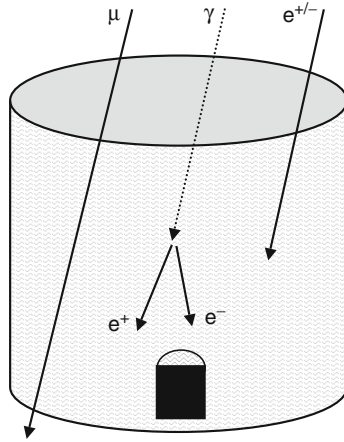
Diagram of a simple muon telescope made of plastic scintillator panels matched to photomultiplier tubes

are necessary to account for showers that are not detected, thereby recovering the rate at the top of the atmosphere. Then, in order to interpret the gamma-ray rate, the measurement must be placed in an energy range. Simulations are necessary to determine the range of energies for gamma rays generating the showers that are detectable by the instrument. Finally, the measured rate of air showers does not represent only gamma rays, but also includes relativistic charged particles capable of generating air showers that reach the ground. Air showers generated by cosmic rays outnumber those from gamma rays by orders of magnitude. The simple array proposed above gives little power to reject this substantial background. The more sophisticated techniques described next address some of these challenges.

4.2 Water Cherenkov Telescopes

Water Cherenkov telescopes (WCTs) detect an air shower when particles generated in the cascade survive to ground level and impact the instrument directly. This has several implications for the telescope design. As with the scintillator arrays discussed above, the array of detectors that make up the telescope must cover a very large area on the ground. This is important both to constrain the size and location of the shower on the ground and to increase the overall collection area and therefore, the sensitivity. Minimally, the telescope should encompass an area large enough to contain the lateral extent of a shower, which depends on the primary energy of the gamma ray and on the altitude of the detector. A 1 TeV gamma ray will produce a shower with a diameter of 100–200 m on average at sea level. If the shower is not laterally contained, then the reconstruction of the direction, the estimate of the energy, and the efficiency for gamma-ray selection are all severely compromised. Detectors that cover area beyond that required to contain the shower allow the collection of more showers per source, or direction on the sky, in the available observation time. Larger detectors also increase the knowledge of the shower size and shape on the ground, allowing more efficient rejection of the cosmic-ray background.

$$\text{Sensitivity} \propto \frac{\text{Signal}}{\sqrt{\text{Background}}} \propto \frac{\text{Area} \times \text{efficiency} \times \text{time}}{\sqrt{\text{Area}}}$$



■ Fig. 11-6

Diagram of a basic water Cherenkov detector consisting of a tank filled with water and housing a photomultiplier tube. Relativistic charged particles, for example, electrons (e^-) and positrons (e^+) passing through the tank generate Cherenkov light in the water until they lose energy and fall below the emission threshold. Secondary gamma-ray emission interacts with the water to produce additional electron-positron pairs that also generate Cherenkov emission. A muon (μ) generated by cosmic-ray interactions in the atmosphere also generates Cherenkov emission in water but can pass through the detector completely

Although the telescope must be large enough to contain and characterize the shower, the individual detectors do not necessarily need to be densely packed. Large showers caused by more energetic primaries that deposit many particles on the ground only need to be sampled sparsely to obtain a good knowledge of the shower characteristics. An example of a very sparse instrument is the ground array portion of AUGER, an ultra-high-energy cosmic-ray detector, which includes 1,600 water Cherenkov detectors (● Fig. 11-6) separated by 1.5 km throughout a 3,000-km² region (Abraham et al. 2010). However, the fractional coverage of an array becomes very important for smaller, lower energy showers that have correspondingly smaller numbers of particles reaching the ground. Higher density arrays of water Cherenkov detectors provide sensitivity to gamma rays below 100 TeV. Closely packed arrays or instrumented pools, such as Milagro (Atkins et al. 2000, 2003), provide effective area comparable to satellite-based instruments near 100 GeV. Further discussion of the current capabilities and future potential for this technique may be found in Buckley et al. (2008).

An obvious but important general feature of particle air shower detectors is that they are not pointed telescopes. The detectors sit in fixed configurations on the ground and observe the sky as it drifts overhead. WCTs and similar instruments are by nature very wide-field instruments. The main limitations on the field of view are the local horizon and the increasing depth of atmosphere that must be crossed by showers at low elevation angles. These instruments naturally observe ~ 2 sr of sky simultaneously and sweep out a full band in right ascension over the course of a sidereal day.

The particles that reach the ground, mostly electrons and positrons in the case of gamma-ray-induced air showers, are still moving at velocities close to the speed of light. Relativistic

charged particles incite lower energy radiation in a medium when the speed of the particle exceeds the speed of light in that medium, an effect known as Cherenkov radiation (Čerenkov 1937). If the medium is transparent to the Cherenkov light, which peaks in the blue segment of the optical band, then optical detectors can be used to detect the presence of the particles. Cherenkov light is generated in a cone along the direction that the particle is traveling. The angle of this cone is determined by the density of the material and the velocity of the particles.

In water, the Cherenkov angle is close to 41° . A light sensor lying within the Cherenkov cone will detect a rapid flash from a charged particle passing nearby. If the light is not highly attenuated by the medium, then a single sensor can cover a large volume. Water is one of the most useful materials for Cherenkov detection because in addition to generally being cheap and easy to work with, it also transmits Cherenkov radiation over a reasonable distance. Attenuation lengths exceeding 10 m are easily achievable. Whereas scintillator detectors only observe particles that pass directly through, a single light sensor within a pool of water can detect air shower particles passing within a 10 m radius of the sensor. This allows for a very high, in fact almost complete, efficiency for detecting particles within the detector area as compared to the few percent fractional area accessible to sparse arrays of scintillator detectors. Note that this also means that the energy threshold for a telescope employing an array of discrete water Cherenkov detectors will have a dependence on the spacing of the detectors.

An array of light sensors distributed throughout a pool of water or among an array of tanks (► Fig. 11-6) provides an image of the charged particles in the shower at ground level. A pool containing an array of sensors or an array of discrete water tanks (or a combination of both) that is large enough to contain the lateral extent of the shower can measure both the size and intensity of the shower front. When a significant number of sensors detect Cherenkov light, the telescope is triggered and the arrival time and amplitude of the light signals from each sensor are read out as a candidate shower. The arrival times of the signals are used to reconstruct the direction of the shower and thus the direction of the incoming gamma ray. The shower front is not the perfect plane idealized in the cartoon of ► Fig. 11-4, but is instead slightly conical. Corrections to the shape of the shower front to allow the best direction to be found depend on knowing the position of the center, or core, of the shower, where the gamma ray, unimpeded, would have struck the ground. This is one reason that a WCT should be large enough to contain the particle content that reaches the ground.

Water also converts secondary energetic photons into charged particles within a couple of meters.⁴ These secondary cascades allow the detection of the energetic photons in an air shower in addition to the charged particles, further increasing the detection efficiency of the instrument and providing a more complete measurement of the energy in the shower at ground level.

Cosmic-ray air showers greatly outnumber gamma-ray air showers and form the dominant background for detecting gamma-ray sources. As discussed earlier, gamma-ray-induced showers consist primarily of electrons, positrons, and high-energy photons. Cosmic-ray nuclei also produce cascading air showers, but these proceed primarily through hadronic interactions that fragment nuclei and produce a broader array of particles including pions, muons, and neutrinos in addition to electromagnetic cascades. The muons generated in these interactions provide a means of distinguishing air showers with hadronic components, but the muons themselves can produce a background for detecting gamma rays. Muons are more penetrating than electrons,

⁴This is one reason that certain nuclear reactors are operated in a pool of water. In addition to cooling, the water moderates the fission process and absorbs radiation. The water glows a distinctive blue from the Cherenkov light produced by relativistic electrons.

and the energetic muons commonly reach ground level. As they pass through the atmosphere or the water in a Cherenkov detector, they generate a distinctive ring of light. In a water Cherenkov detector, a muon can be distinguished because it passes through the detector, beyond the interaction depth for electrons. The presence of muons in a shower provides one key diagnostic for rejecting air showers produced by cosmic rays. However, muons passing through a pool of water may be viewed by several light sensors and will trigger the data acquisition unnecessarily. For this reason, it is beneficial to optically isolate the sensors used in a water Cherenkov telescope. The isolation prevents a single particle from contributing light to many sensors and mimicking a shower.

The high levels of background generated by cosmic rays create substantial challenges for operating a water Cherenkov telescope. The high rate of events is difficult to remove without using information about the shower direction and distribution available in high-level data reduction. For example, the plane of the shower front and the position of the center of the shower on the ground impact the ability to discern a cosmic-ray-generated air shower from a gamma-ray-generated one, but these characteristics are not known until after the data is analyzed. In other words, these detectors trigger and read out thousands of light flashes per second, and most are background to the desired gamma-ray signals. In order to not miss recording the gamma-ray showers, fast timing and low dead time of the digitizing electronics must be paired with good computing capabilities located at the telescope. The data can be processed in real time to allow the best gamma-ray candidates to be read out and recorded despite the continuous onslaught of showers. This is particularly important to enable the observation of more small showers caused by lower energy gamma rays. Even after reconstructing the shower properties, water Cherenkov data is still dominated by cosmic-ray showers.


The rate of cosmic rays at the top of the atmosphere is essentially uniform from all directions in the sky. In order to quantify the background recorded by a water Cherenkov telescope, the data can be mapped in detector coordinates to reveal the background acceptance of the instrument. The measured cosmic-ray acceptance provides a template to model the background contribution to any part of the celestial sky over a specified time period. Strong divergences from this background template indicate the positions of gamma-ray sources and allow the calculation of a significance map for gamma rays, giving the probability of a gamma-ray source at each location. This is different from using background subtraction, but very useful and necessary in the case where the background rejection is limited and gamma rays can only be confidently identified statistically, not on an event-by-event basis.



There are several interesting situations where the background limitation of water Cherenkov telescopes is largely mitigated. The constant rate of the background means that a sufficiently brief analysis, such as the few second to minute-long duration applicable to gamma-ray bursts, is no longer background dominated but instead signal limited. WCTs have excellent sensitivity to short events, and if a GRB emits enough photons above 100 GeV, it should be easily detected in the field of a WCT with a suitably low energy threshold. The background spectrum falls as a power law with energy, just as the cosmic-ray spectrum does. This also means that sources with spectra that fall less quickly have an advantage at higher energy. The coverage area and integration times achievable with WCTs, both of which increase sensitivity, and the decreasing background rate with energy provide some advantages for the study of the highest energies of very-high energy sources with these instruments. In fact, several of the candidate sources identified first by the Milagro instrument have been this type of object (Abdo et al. 2007).

There are several very difficult backgrounds for detecting gamma-ray air showers even in the case of a perfect reconstruction knowledge on the ground. A primary proton can interact to

produce a neutral pion, which then decays to gamma rays. The air shower initiated by the secondary gamma rays is purely electromagnetic and therefore practically indistinguishable from a gamma-ray-induced shower. Cosmic-ray electrons also generate showers that are nearly identical to those formed by gamma rays with the exceptions of a small difference in the depth of the shower maximum and the additional presence of Cherenkov light caused by the primary electron. Both of these backgrounds are relatively small compared to the substantial population caused by more common proton interactions and do not limit current telescopes. However, the electron background gains prominence quickly when attempting to lower the energy threshold below 0.1 TeV.

4.3 WCTs: A Little History

Large arrays of scintillator detectors have been used very successfully to detect extensive air showers generated by ultra-high-energy cosmic rays. Although such arrays are capable of detecting showers from extremely energetic gamma rays, no definitive detections of gamma-ray sources have been made using sparse arrays (Alexandreas et al. 1992; Borione et al. 1994). This simply suggests that there are not many gamma rays to detect at such high energies. The sensitivity of scintillator arrays is limited by the difficulties of detecting gamma rays in the energy range where they are more common, below ~ 100 TeV. At lower gamma ray energies, the particle density in an air shower at ground level is low, and distinguishing air showers generated by gamma rays from those generated by far more plentiful cosmic rays becomes very difficult. This is where the properties of water Cherenkov detectors provide a significant advantage. Water Cherenkov detectors were used to study cosmic-ray showers in the 1950s (Watson 2011), and they currently play an important role in the detection of ultra-high-energy cosmic rays (Abraham et al. 2010). However, the first instrument to successfully apply the water Cherenkov technique to the study of gamma ray sources was the Milagro telescope (Atkins et al. 2003) shown in  Fig. 11-7. The Milagro prototype first detected VHE gamma rays from the active galactic nuclei, Mrk 501, during a flaring episode in 1997 (Atkins et al. 1999). By 2008, the completed Milagro instrument found signals of TeV gamma rays coming from a variety of distinct objects that included blazar galaxies, Galactic pulsar wind nebulae, and other sources with undetermined origins (Abdo et al. 2007). Milagro also mapped the spatial distribution of VHE gamma rays from the Galaxy and found an unexplained excess from the Cygnus Arm (Abdo et al. 2008).

A successor to the Milagro Observatory, the High-Altitude Water Cherenkov Experiment (HAWC) is now under construction by an international collaboration at a high-altitude site (~ 4 km) at Sierra Negra in Mexico. The new site and planned configuration of large, isolated tanks ( Figs. 11-8 and  11-9) will allow substantial improvements in sensitivity of the technique and the angular resolution for gamma rays.

4.4 Imaging Atmospheric Cherenkov Telescopes

The relativistic charged particles in an extensive air shower generate Cherenkov light through interactions with molecules in the atmosphere. The Cherenkov light is at optical wavelengths and transmits through air without significant absorption. The light is very faint compared to sunlight and moonlight and is not visible without a telescope. The signal is very short in duration, but a sufficiently fast sensor enables detection above the background light if the sky is



■ Fig. 11-7

An aerial view of the Milagro gamma ray observatory near Los Alamos, New Mexico, USA (Courtesy of the Milagro Collaboration <http://umdgrb.umd.edu/>)



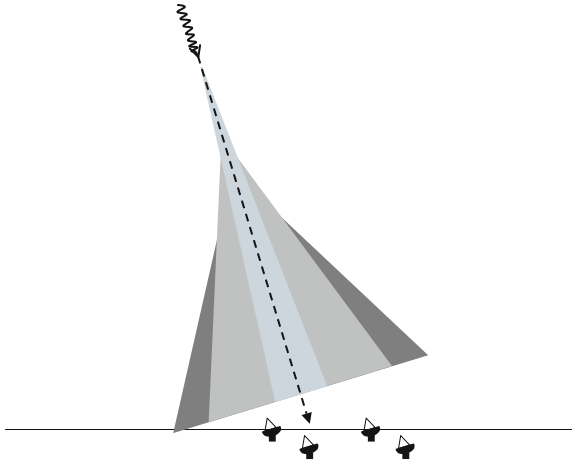
■ Fig. 11-8

The planned layout for the High-Altitude Water Cherenkov (HAWC) Observatory, an array of 300 water Cherenkov tanks at Sierra Negra in Mexico. Existing prototype tanks are visible at the *left side* of the photo. The designed configuration is overlaid (Courtesy of the HAWC Collaboration <http://hawc.umd.edu/>)



■ Fig. 11-9

One of the 7.2×4.3 -m tanks that will be used in the HAWC observatory (Courtesy the HAWC Collaboration <http://hawc.umd.edu/>)



■ Fig. 11-10

Cartoon of the Cherenkov light generated by an extensive air shower passing through the atmosphere. The shower core dominates the emission. The angle of the cone of Cherenkov light depends on the density of the air and therefore evolves with altitude

reasonably dark. The Cherenkov emission allows the measurement of an air shower at a significant distance using an optical image (see ● Fig. 11-10 to get an idea of this). The numerous charged particles in a shower create a pool of light on the ground. A telescope sitting anywhere within the Cherenkov pool can record an image of the shower cascading through the sky overhead. Even a small telescope has a large effective size because the ground covered by the Cherenkov pool has a diameter of about 200 m at sea level.

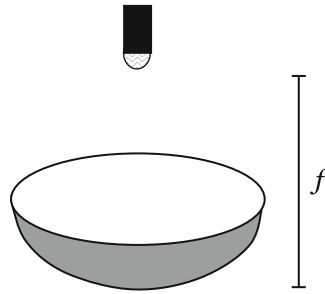


Fig. 11-11

Diagram of a very simple Cherenkov telescope made of a spherical reflector with a photomultiplier tube placed at the focal point. Additional phototubes and reflectors allow simple rejection of background light by requiring a temporal coincidence of signals from the same shower

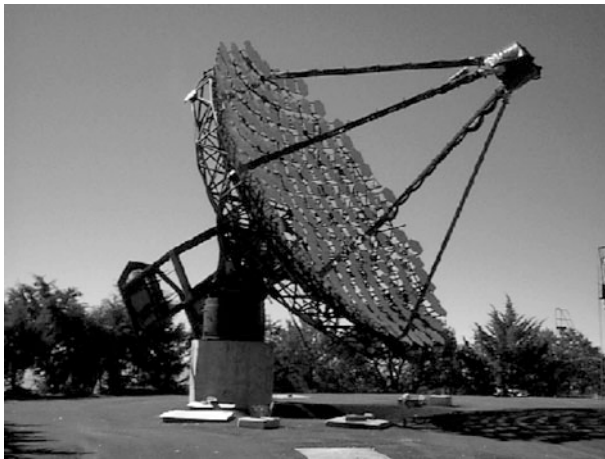
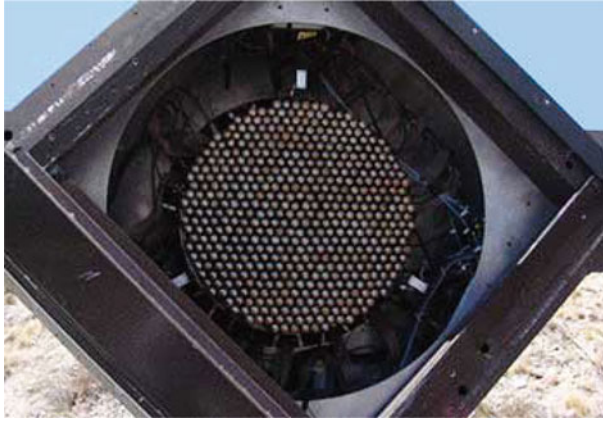


Fig. 11-12

The Whipple 10 m reflector located at Mt. Hopkins in Arizona was the first IACT to make a significant detection of a TeV gamma-ray source (Courtesy the VERITAS Collaboration <http://veritas.sao.arizona.edu/>)

Cherenkov signals from air showers can be detected using a fast photo-sensor, such as a photo-multiplier tube, pointed at a dark patch of sky. However, focusing systems are necessary to constrain the shower properties and to reach a useful sensitivity level (► Fig. 11-11). Moderately large collection dishes, of order 10 m allowing more than 100 m² in reflector area (► Fig. 11-12), are favored because larger reflector area enhances sensitivity to lower energy showers. Put simply, lower energy showers produce less particles and therefore less Cherenkov photons. The lower the intensity is on the ground, the more important it is to collect photons over a larger area in order to pass the minimum threshold to image the shower. Near the threshold energy of the telescope, the reflector area is one of the most important factors affecting the sensitivity. However, the benefits of larger structures are offset by their costs and complexities.



■ Fig. 11-13

A Cherenkov camera used in one of the VERITAS telescopes is visible in a mounting box (1.8 m sides). The camera views a 3.5° wide region of sky with 499 pixels spaced 0.15° apart (Courtesy the VERITAS Collaboration <http://veritas.sao.arizona.edu/>)

Beyond the collection area, the optical requirements are not very strict. The mirror area should be large but can be segmented. Many Cherenkov dishes are based on the Davies-Cotton design, which specifies segmented mirrors on a spherical or parabolic primary, developed for solar concentrators (Davies and Cotton 1957). This configuration allows some image distortions that increase with the angle from the optical axis. This limits the field of view, but a reasonable amount of aberration is tolerable given the low resolution necessary for adequate measurement of the Cherenkov signal. Distortions can be safely ignored if their size can be contained within a single camera pixel. Current instruments favor pixels with diameters of about 0.2° . The parabolic mirror configuration also introduces temporal distortion, a spreading in the arrival times of photons at the camera, but this is unimportant as long as the spread in the time is smaller than the intrinsic width of the Cherenkov shower front, ~ 5 ns.

A less important consideration for the design of support structures for imaging Cherenkov telescopes is the ability to rapidly point the reflector toward a specified part of the sky. Gamma-ray bursts have not yet been detected by ground-based telescopes, but their emission is expected to decay quickly after the onset of the burst, perhaps lasting only a few seconds to minutes. This means that even at a very fast slew rate of 3° s^{-1} , an event happening in the opposite direction in the sky cannot be viewed for about half a minute. Additionally, as Cherenkov telescopes increase in sensitivity and devote less exposure for each observation of a source, the time spent between targets begins to make a more noticeable impact on the total amount of observing time that can be obtained. The total time available in a year may only be 800–1,000 h. The faintness of the Cherenkov signal requires darkness and limits high-quality observations to times when the Moon is far from the field being observed.

The Cherenkov signal is very rapid. It lasts for approximately 5 ns for 1 TeV showers at vertical incidence and near the core of the shower. To match this signal size and to help in separating shower signals from background light, a Cherenkov camera (e.g., ► Fig. 11-13) must make very fast images. Current state-of-the-art cameras produce signal pulses with rise times of a few ns. These are well-matched to fast electronic systems, which sample signals at rates around 1 GHz

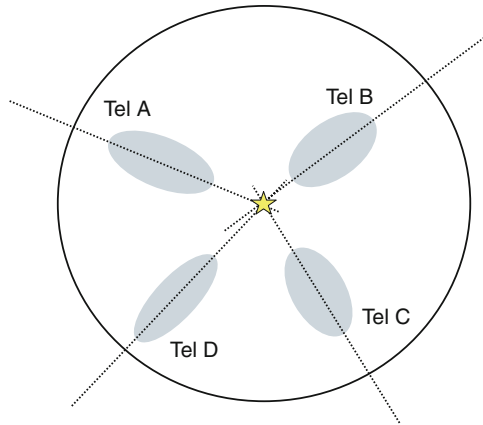
(1 ns frames). The fast sampling is important to reduce the confusion of Cherenkov signals from air showers with the continuous bombardment of background photons and muon signals. Because the intensity of the light is an important gauge of the energy in the shower and the major limitation at low energy, camera pixels must be able to characterize the signal generated at the single photon level. Cherenkov cameras are used to count every photon that arrives at the image plane.

A camera should have suitable pixelation to resolve the shower image and constrain broad structural features. The typical height of shower maximum at 1 TeV is about 10 km, meaning that the image of the shower extends over about 1° of sky. The core of the shower makes up a fraction of this, setting a basic scale for the imaging at $\sim 0.1^\circ$. Pixels covering 1° are sufficient for detecting showers, particularly those with energies above 1 TeV, but pixels closer to 0.1° offer several advantages. Smaller pixels allow improved fits of the axis of the ellipse and therefore better resolution for the direction of the gamma ray. Increased resolution on the structures in the shower allows better rejection of non-gamma-ray air showers. Reduced pixel area also helps with limiting the ever-present noise background from stars and scattered light in the atmosphere. In other words, the smaller the area of sky covered by a pixel, the lower the amount of background light level collected in that pixel. The benefits of pixels with diameters less than $\sim 0.1^\circ$ are outweighed by the cost and complexity of the increased number.

The size of the air showers also set a scale for the field of view that the camera should cover. In order to contain the full image, a requirement important for measuring both the direction and energy content of the shower, the camera should view several degrees. Particularly for higher energy showers at larger offsets, larger cameras are necessary to contain the full shower image. Considerations must also be made for analysis and astrophysical implications of the camera size. If a background estimate is made using portions of the sky viewed simultaneously with the source of interest, then the field of view must be large enough to contain both. Gamma-ray sources themselves also provide a minimum guideline. Several prominent Galactic supernova remnants, for example, have diameters of around 1° . Cameras that extend over 5° or more can contain these sources and allow a well-separated background measurement. Although there are scientific advantages to observing larger fields of view, particularly for surveys, extended structures, and time-domain astrophysics, Cherenkov cameras are limited in size by items including the optical properties of the dish, the supporting mechanical structure, and the expense of pixels and accompanying electronics.

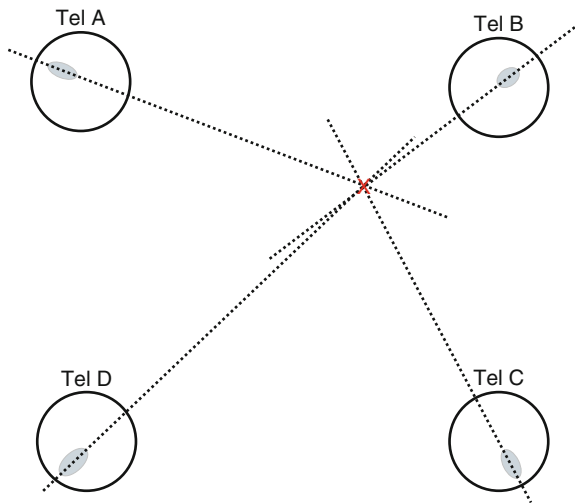
The Cherenkov image of an air shower represents a two-dimensional projection of the passage of the cascade through the atmosphere and can be used to characterize the direction and energy of the incoming gamma ray. The image intensity is dominated by the region of highest particle density along the shower track, the shower core, which appears as an ellipse in the image plane. The eccentricity of the ellipse depends on the distance of the shower track from the optical axis of the telescope on the ground. When the telescope is looking directly along a shower track, the ellipse becomes a circle and the origin of the track becomes ambiguous. When the track is offset, a line drawn through the major axis of the elliptical image crosses the direction of the gamma ray in the sky. For this reason, showers with an axis passing directly through the telescope optical axis are very poorly reconstructed. The best measurements are for showers that are off the telescope axis, but not so far as to be truncated by the edge of the field of view (► [Fig. 11-14](#)).

The image of an air shower is best characterized by knowing not only the direction of the initial gamma ray but also by knowing the location where the shower axis impacts the ground. This fully constrains the path of the air shower through the atmosphere. The geometry of the



■ Fig. 11-14

The image of an air shower in Cherenkov light can be parameterized as an ellipse. The major axis aligns with the direction of the initial gamma ray. Multiple images of the same shower from different telescopes in an array provide constrain to the position of the gamma-ray source on the sky



■ Fig. 11-15

The major axis of the Cherenkov image of an air shower points to the location where the shower core impacts the ground (marked by a red X). Observations by multiple telescopes in an array constrain the location of the core

air shower and the distance that it falls from the telescope allow proper interpretation of the shape and intensity of the camera image. A well-formed image in a single telescope provides some information about the impact distance through the eccentricity of the ellipse, but the measurement is greatly improved with the use of multiple telescopes (👁 Fig. 11-15).

More energetic gamma rays give rise to more secondary particles and thus, more Cherenkov light. In fact, the energy of the shower is proportional to the number of photons in the shower image, often referred to as the *image size*. The number of photons associated with a shower image in the camera indicates a lower limit and must be corrected for the loss of light in traversing the atmosphere and inefficiencies in collection by the telescope mirror and camera. Not much can be done about losses in the atmosphere, besides placing telescopes at higher altitudes to reduce the amount crossed. Losses in the telescope can be minimized by using and maintaining good mirror reflectivity, reducing any gaps in the active surface of the camera, and using photosensors that are tuned to the spectrum of the Cherenkov light, which actually peaks near the ultraviolet. However, the primary limitation for detecting Cherenkov photons comes from the quantum efficiency of the photosensors. Commonly used photomultiplier tubes are only about 20–30% efficient in collecting the photons that strike them. More efficient devices are being explored and provide the most obvious method of obtaining significant increases in efficiency. Although of some importance for the precision of angular and energy measurements of smaller showers, the improvements translate directly into reduced energy thresholds because the faintest showers, those from lower energy gamma-rays, are impacted by the loss of light.

Unfortunately, increasing the sensitivity of these instruments is not as simple as presented above. While signal enhancements are important, the dominant sensitivity limitations come from the backgrounds. IACTs, like WCTs, record thousands of candidate events per second, but only a tiny fraction of these are gamma rays. In order to detect a significant gamma-ray signal, several backgrounds must be rejected at the trigger level, removed from the recorded data, or modeled in the analysis.

IACTs commonly employ a hierarchy of triggers to reject backgrounds and select gamma-ray-like events to be recorded for further analysis. The signal in a single camera pixel must reach the level expected from a photon in order to be considered as part of a shower image. At the camera level, a cluster of neighboring pixels are required for the telescope to register an event. In an array of IACTs, coincident signals exceeding the threshold for a camera are required from multiple telescopes. At each level, the threshold conditions are adjusted to reduce the background contamination and throttle the rate of recorded events to something manageable at the cost of rejecting a minimal level of gamma rays. The gamma rays that are lost are typically those with lower energies, near the threshold for the instrument. Optimal settings for the trigger are important for achieving a low threshold and keeping the amount and rate of recorded data at manageable levels. Further event selection to enhance the signal to noise ratio can be done in later analysis of the recorded data.

Reduction of the threshold requirement for accepting a recorded event in the telescope to the level of a single photon in a single pixel, produces an explosion in the rate of collected data. This is where the threshold is low enough to allow the general population of photons in the night sky to trigger the telescope. The night sky background produces uncorrelated signals in pixels. It can be dramatically reduced by requiring a signal to appear simultaneously in neighboring pixels. For this reason, camera pixels should be small enough that the probability of a chance coincidence of night sky photons in a single pixel is low. Note that this also means darkness of the telescope site is an important factor for detector performance at the low end of the energy range.

Stars make up a foreground for observations by IACTs. A star that shines brightly in the blue band (recall that IACT cameras are designed to be sensitive near the peak frequency for Cherenkov emission) causes a very high rate of triggers in a pixel or could in some cases damage the photosensor. Often stars produce correlated signals in neighboring pixels since the size of a point source in the optics is typically close to the size of a pixel. This means that the image

of a bright star will pass the trigger criteria for a cluster of pixels in a telescope. It will also pass a requirement for multiple telescopes. The locations of bright stars are well known and can be managed at the analysis level. To reduce the rate of unwanted events, the only practical solution is to prevent pixels affected by a bright star from participating in the telescope trigger. The photosensors used in Cherenkov telescopes are by necessity extremely sensitive. In the case where a star is very bright, it may be necessary to disable pixels affected by the star to protect them from damage. In either the case of disabling pixels or masking them out of analysis, stars form a highly problematic structured foreground for gamma-ray observations. As the telescopes further increase in sensitivity and observations of the sky deepen, this becomes more of a limiting factor.

Cosmic rays generate air showers from all directions in the sky at a high rate in the VHE band and drown out gamma-ray sources. Some of the air showers generated by cosmic-ray nuclei look different enough to be separated out based on their image parameters. As discussed in [Sect. 2](#), air showers produced by nuclei have several key differences from those produced by gamma rays. Nuclei most commonly initiate hadronic cascades that produce a less uniform energy deposit over the shower. This creates Cherenkov images that are less uniform along the track, or clumpy, and more laterally broadened. Some nuclei interact to produce a neutral pion that then decays into gamma rays. The following cascade is electromagnetic just like those of primary gamma rays. Although there are subtle differences – the interaction depth for nuclei differs from that for photon pair production and the nucleus generates additional Cherenkov light before the air shower commences – for practical purposes, this population of showers forms a nearly irreducible background for IACT observations. Because the cosmic-ray rates are essentially uniform in direction at the top of the atmosphere, background showers that cannot be rejected can be modeled. This can be done either by taking an observation of a blank field as a template or by using regions without gamma-ray sources within an observed field. Backgrounds that cannot be filtered out must be modeled by building a predicted profile for detected cosmic-ray showers over an observed field.

The range of energies accessible to this type of telescope is limited at low energy by the ability of the telescope to collect fainter Cherenkov signals and distinguish them from backgrounds. Even if a shower is detectable below a few hundred GeV, there is little information available to determine whether the origin is a gamma ray, a cosmic ray, or a single muon passing near the telescope. Muons passing close enough to a telescope produce a distinctive ring image, but the images can pass simple trigger requirements at the telescope level, especially in cases where only a portion of the ring is visible in the camera and appears similar to a small shower image. The rate of muons, like the rate of their cosmic-ray parents, is much higher than that for gamma rays and provides a severe limitation to the minimum energy that can be reached by a single IACT. A simple and powerful method of reducing muon contamination is to require a coincidence among multiple telescopes. While an air shower will appear in all telescopes within the $\sim 10^5$ -m² light pool, a muon signal only appears in the telescope within range of the muon, some 10s of m in the impact parameter. This also sets a practical lower limit for array spacing.

At high energy, observations are statistically limited. The gamma-ray and cosmic-ray air showers are more easily distinguished with increasing energy, but signal events become increasingly rare. The flux of very-high-energy gamma-ray sources even in optimistic cases falls as E^{-2} or more. The most obvious way to enhance the sensitivity to higher energy events is to cover more area on the ground.

Coordinated arrays of IACTs can cover very large areas. Simply adding telescopes increases the number of showers visible to the instrument and reduces some backgrounds. A little tuning

of the placement has substantial benefits for reconstructing the gamma-ray direction and energy and for reducing backgrounds. When telescopes view the same field, events local to a single telescope, often muons, can be easily rejected. Two or three telescopes within about 100 m of one another are close enough to fall within the Cherenkov light pool. Multiple images of the same shower constrain the direction to improve angular resolution and the impact location of the shower core on the ground to improve the energy measurement.

The sensitivity of an array of imaging Cherenkov telescopes depends strongly on the size of the telescopes in the array, the area of ground that the array covers, and the layout of telescopes within the array. Larger telescopes provide increased telescope mirror area that enhances the detection of faint showers by collecting more Cherenkov photons per shower image (e.g., ● Fig. 11-16). Faint showers are produced by lower energy gamma rays, and so larger dishes translate into more sensitivity at low energies and a reduction in the energy threshold. Increasing the area of ground covered by the array increases the number of air showers that can be detected. Observations at higher energy are more statistically limited, and so larger arrays translate into better sensitivity to higher energy gamma rays and an increase in the upper energy range. In the signal-dominated regime, the sensitivity increases proportional to the area. The arrangement of the telescopes in the array, primarily characterized by the separation distance between telescopes, has a secondary influence on the sensitivity. Closer telescope spacings favor smaller, fainter, lower energy showers, while large spacings favor larger, higher energy showers.



■ Fig. 11-16

The central, very large telescope structure is a 30-m addition to the center of the HESS array of 12-m dishes. It is designed to lower the energy threshold for the array below 100 GeV (Courtesy the HESS collaboration <http://www.mpi-hd.mpg.de/hfm/HESS/>)

4.5 IACTs: A Little History

Ground-based gamma-ray astronomy has become a thriving field containing multiple facilities and exciting results that are stimulating work in a variety of areas. The current successes owe a great deal to the vision, effort, and substantial persistence of early pioneers. It took nearly half a century to achieve the sensitivity to detect the first sources, long after the promise of the technique was recognized. Galbraith and Jelley (1953) reported the successful detection of cosmic-ray air showers as early as the 1950s. In the 1960s Chudakov and Zatsepin in Crimea and Jelley and Porter in the United Kingdom pursued observations of showers from gamma rays. Weekes continued the work of Porter and Jelley, working with the Whipple collaboration in Arizona to obtain the first high significance detection of the Crab Nebula using a 10-m reflector at the Fred Lawrence Whipple Observatory in 1989 (Weekes et al. 1989). Weekes (2005) reviews the development of the technique.

The Whipple team continued to innovate camera and telescope designs and followed the discovery of the Crab with discoveries of several dramatically variable TeV blazars. The HEGRA telescopes made the first successful observations using an array of small IACTs (Pühlhofer et al. 2003). These efforts and their successes launched a second generation of instruments around the world. These facilities have explored enhancements to further refine the technique, including higher efficiency cameras, faster and smarter triggering and digitization, wide-field optics, array configurations capable of extending the energy range, and photosensors and methods for operating under moonlight. The H.E.S.S. collaboration operated the first array of medium-class reflectors (~12-m-diameter dishes) (Bernlöhr et al. 2003), producing the first map of a resolved VHE gamma-ray source (Aharonian et al. 2006a) and revealing dozens of new sources in a survey of the central portion of the Galactic plane (Aharonian et al. 2006b). The MAGIC collaboration constructed the first large dish (17-m-diameter frames) to enhance sensitivity below 100 GeV and has also explored high efficiency photo-sensors and active mirror control (Baixeras et al. 2004). The MAGIC team exploited these capabilities to detect the first pulsar signal using a ground-based gamma-ray telescope, finding the 33 ms period of the Crab Pulsar (Aleksić et al. 2011) (clearly not wanting to be forgotten as the queen source of TeV gamma-ray astronomy). The VERITAS observatory, a four-telescope array (Holder et al. 2006), has since extended this work to detect pulsations above 100 GeV (Aliu et al. 2011). For a recent review of TeV astronomy, see Hinton and Hofmann (2009).

An international consortium plans to build a next-generation observatory, the Cherenkov Telescope Array (CTA) (The CTA Consortium 2010). The facility will be a much larger array of IACTs than current instruments and may include various enhancements to the optics, cameras, and electronics that will increase sensitivity by over an order of magnitude and further extend the energy reach of the technique.

References

- | | |
|---|---|
| Abdo, A. A., Allen, B., & Berley, D., et al. 2007, <i>ApJ</i> , 664, L91 | Abdo, A. A., Ackermann, M., & Ajello, M., et al. 2010, <i>A&A</i> , 523, L2 |
| Abdo, A. A., Allen, B., & Aune, T., et al. 2008, <i>ApJ</i> , 688, 1078 | Abdo, A. A., Ackermann, M., & Ajello, M., et al. 2011, <i>Science</i> , 331, 739 |
| Abdo, A. A., Ackermann, M., & Ajello, M., et al. 2009, <i>Nature</i> , 462, 331 | Abraham, J., Abreu, P., & Aglietta, M., et al. 2010, <i>Nucl. Instrum. Methods Phys. Res. A</i> , 613, 29 |

- Aharonian, F., Akhperjanian, A. G., & Bazer-Bachi, A. R., et al. 2006a, *A&A*, 449, 223
- Aharonian, F., Akhperjanian, A. G., & Bazer-Bachi, A. R., et al. 2006b, *ApJ*, 636, 777
- Aharonian, F., Akhperjanian, A. G., & Bazer-Bachi, A. R., et al. 2007, *A&A*, 464, 235
- Aleksić, J., Alvarez, E. A., & Antonelli, L. A., et al. 2011, *ApJ*, 742, 43
- Alexandreas, D. E., Allen, R. C., & Biller, S. D., et al. 1992, *Nucl. Instrum. Methods Phys. Res. A*, 311, 350
- Aliu, E., Arlen, T., & Aune, T., et al. 2011, *Science*, 334, 69
- Atkins, R., Benbow, W., & Berley, D., et al. 1999, *ApJ*, 525, L25
- Atkins, R., Benbow, W., & Berley, D., et al. 2000, *Nucl. Instrum. Methods*, A449, 478
- Atkins, R., Benbow, W., & Berley, D., et al. 2003, *ApJ*, 595, 803
- Atwood, W. B., Abdo, A. A., & Ackermann, M., et al. 2009, *ApJ*, 697, 1071
- Baixeras, C., Bastieri, D., & Bigongiari, C., et al. 2004, *Nucl. Instrum. Methods Phys. Res. A*, 518, 188
- Berger, M. J., Hubbell, J. H., & Seltzer, S. M., et al. 2010, XCOM: Photon Cross Sections Database, (version 1.5), [Online] Available: <http://www.nist.gov/xcom>
- Beringer, J., & (Particle Data Group) et al. 2012, *Phys. Rev.*, D86, 010001 (<http://pdg.lbl.gov>)
- Bernlöhr, K., Carrol, O., & Cornils, R., et al. 2003, *Astropart. Phys.*, 20, 111
- Borione, A., Covault, C. E., & Cronin, J. W., et al. 1994, *Nucl. Instrum. Methods Phys. Res. A*, 346, 329
- Buckley, J., Byrum, K., Dingus, B., et al. 2008, arXiv:0810.0444 [astro-ph]
- Čerenkov, P. A. 1937, *Phys. Rev.*, 52, 378
- Davies, J. M., & Cotton, E. S. 1957, *Sol. Energy*, 1, 16. The Proceedings of the Solar Furnace Symposium
- Esposito, J. A., Bertsch, D. L., & Chen, A. W., et al. 1999, *ApJS*, 123, 203
- Gaisser, T. K. 1991, *Cosmic Rays and Particle Physics* (Cambridge: Cambridge University Press)
- Galbraith, W., & Jelley, J. V. 1953, *Nature*, 171, 349
- Hinton, J. A., & Hofmann, W. 2009, *ARA&A*, 47, 523
- Holder, J., Atkins, R. W., & Badran, H. M., et al. 2006, *Astropart. Phys.*, 25, 391
- Pühlhofer, G., Bolz, O., & Götting, N., et al. 2003, *Astropart. Phys.*, 20, 267
- Rossi, B., & Greisen, K. 1941, *Rev. Mod. Phys.*, 13, 240
- Schoenfelder, V., Aarts, H., & Bennett, K., et al. 1993, *ApJS*, 86, 657
- Su, M., Slatyer, T. R., & Finkbeiner, D. P. 2010, *ApJ*, 724, 1044
- Tavani, M., Barbiellini, G., & Argan, A., et al. 2009, *A&A*, 502, 995
- The CTA Consortium. 2010, arXiv:1008.3703
- Thompson, D. J., Bertsch, D. L., & Fichtel, C. E., et al. 1993, *ApJS*, 86, 629
- von Ballmoos, P., Diehl, R., & Schoenfelder, V. 1989, *A&A*, 221, 396
- Watson, A. A. 2011, *Nucl. Phys. B Proc. Suppl.*, 212, 13
- Weekes, T. C. 2005, Lectures given at the International Heraeus Summer School, eprint arXiv:astro-ph/0508253
- Weekes, T. C., Cawley, M. F., & Fegan, D. J., et al. 1989, *ApJ*, 342, 379

12 Instrumentation and Detectors

Ian S. McLean · James Larkin · Michael Fitzgerald

Department of Physics and Astronomy, University of California,
Los Angeles, CA, USA

1	<i>Introduction</i>	508
2	<i>Classification of Instruments</i>	509
2.1	Camera Systems	509
2.2	Coronagraphs	511
2.3	Spectrometer Systems	512
2.4	Integral Field Spectrometers	517
2.5	Polarimeter Systems	520
2.6	Interferometers	521
3	<i>Detectors and Materials</i>	523
3.1	Classification of Detectors	523
3.2	Semiconductors	525
3.3	Photoconductors	526
3.4	Photodiodes	527
3.5	Applications to CCDs and IR Arrays	528
3.6	Detectors for High Energy	532
3.7	Thermal Detectors	533
3.8	Coherent Detectors	533
4	<i>Cryogenics and Vacuum Systems</i>	535
	<i>References</i>	538

Abstract: This chapter contains a broad introduction to astronomical instruments and detectors. The basic design principles for cameras, spectrometers, polarimeters, and interferometers are given, together with some practical material on instrument building techniques, including vacuum-cryogenic methods. Different detector technologies are introduced, such as CCDs and infrared arrays, together with basic information on semiconductors.

Keywords: Cameras, Coronagraphs, Detectors, Interferometers, Polarimeters, Spectrometers

List of Abbreviations: *AO*, Adaptive optics; *CCD*, Charge-coupled device; *IF*, Intermediate frequency; *IFS*, Integral Field Spectrometer; *IR*, Infrared; *LO*, Local oscillator; *MCP*, Microchannel plate; *MOS*, Metal oxide semiconductor; *NASA*, National Aeronautics and Space Administration; *PMT*, Photomultiplier tube; *PSF*, Point spread function; *SIS*, Superconductor Insulator Superconductor; *UV/O/IR*, Ultraviolet/optical/infrared; *VLT*, Very large telescope; *VPH*, Volume Phase Holographic

1 Introduction

When viewed across the electromagnetic spectrum, astronomical instrumentation can appear quite different, in part because of the wide range of technologies involved, but variety is fundamentally limited by the properties that can be measured. Following the scheme offered in McLean (2008), most instruments can be placed in one of four categories: a photometer or radiometer for measuring the brightness and direction of radiation; a spectrometer for measuring the distribution of brightness as a function of wavelength, frequency, or energy; a polarimeter to determine the degree of alignment or handedness of wave vibrations; or an interferometer which relies on coherent phase relationships to achieve interference effects prior to detection.

All instruments contain some kind of radiation detector and, once again, the range of detector technologies is broad. For most applications in astronomy, however, there are three main classifications of detectors as follows. Photon detectors in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma rays to the far-infrared. Thermal detectors in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical conductivity; thermal detectors have a broad spectral response but are often used for infrared and submillimeter detection. Coherent detectors in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

Breakthrough discoveries in astronomy often rely on the development of new technologies. Among the technologies that have made a difference, both on the ground and in space, in recent years are larger and more sensitive charge-coupled devices (CCDs) and infrared array detectors; more powerful cameras and spectrometers; improved methods for building very large optical/IR telescopes and for building efficient survey telescopes; advances in optics and detectors for X-ray astronomy; devices for the study of the cosmic microwave background; new digital signal processing techniques and new receiver/antenna designs for radio astronomy. Especially impressive has been the advent of adaptive optics techniques that enable large ground-based telescopes to operate at their ultimate diffraction limit.


When the silicon CCD was introduced into astronomy in 1974 (see Janesick 2001), it completely revolutionized astronomical imaging. From a modest 10,000 pixels in the early days, individual CCDs are now 4–16 million pixels, and many instruments employ large mosaics of CCDs to push the total number of pixels up to about one billion. A remarkable feature of the CCD is that it can also detect X-ray photons, and so CCDs are found in the X-ray cameras on the Chandra X-ray Observatory (Garmire et al. 2003). Modern CCDs can now be used for ultraviolet work, but competing devices such as the microchannel plate (MCP) have been used successfully on space missions such as GALEX (Morrissey et al. 2005). Silicon's band gap restricts the use of CCDs to wavelengths less than 1,100 nm, and thus a different kind of detector is needed for the infrared. Usually referred to as infrared arrays, these devices do not employ the charge-coupling principle, and semiconductor materials with a lower band gap must be used. In the near-infrared, these array detectors are now at the 4–16 megapixel size (McLean 2008). For even longer wavelengths, devices known as Transition Edge Sensors (TES) that rely on superconductivity have been developed into arrays for submillimeter astronomy. At the other end of the spectrum, high-energy astronomers have obtained images using large pixel arrays of cadmium zinc telluride (CZT) to detect gamma rays. One example is the Burst Alert Telescope (BAT) on NASA's *Swift* satellite which has 32,768 CZT detectors covering a focal plane of 1.2×0.6 m (Gehrels et al. 2004).

As telescopes have grown in size and instruments have become more costly, a strong trend toward general user facilities and large collaborations has occurred. Most astronomical instruments are therefore highly automated and capable of remote operation. Inevitably, these instruments must be well engineered for reliability. Many important factors constrain the design of new astronomical instrumentation. Engineering details are beyond the scope of this article, but the following sections will serve as a starting point. Clearly, the choice of instrument and the details of the design must depend on the science to be done. This chapter contains an introduction to astronomical detectors and provides basic design principles for cameras, spectrometers, polarimeters, and interferometers. Also included is some material on instrument building, including vacuum-cryogenic techniques. More details can be found in other chapters in this volume and in the references given.

2 Classification of Instruments

Astronomical instruments are often grouped into four classes: (1) photometers/cameras, (2) spectrometers, (3) polarimeters, and (4) interferometers. Although the methods of implementation differ considerably, variations of these instruments exist from X-ray wavelengths to radio wavelengths. The descriptions which follow are mainly applicable for UV, visible, and infrared wavelengths (UV/O/IR).

2.1 Camera Systems

In the simplest camera, the detector (CCD or other array device) is placed directly in the focal plane of the telescope behind a light-tight shutter. Filters to select different wavelength bands are therefore located in the converging beam from the telescope. An alternative approach, shown in  Fig. 12-1, is to collimate the beam by placing a lens after the focal plane at a distance s equal

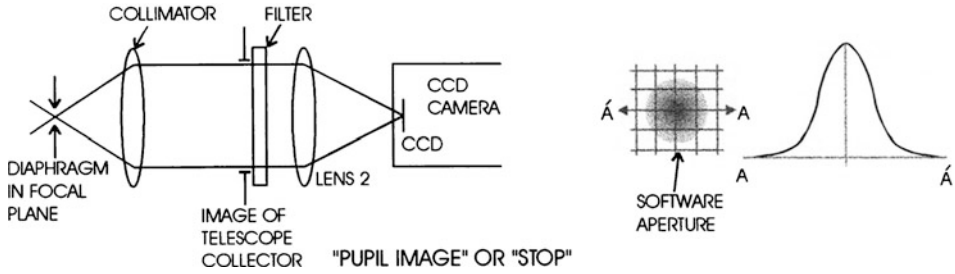


Fig. 12-1

The layout for a basic camera system in which optics are used to collimate the diverging beam from the telescope focal plane and reimage the field at a different magnification. Photometry is performed using software apertures on the digital image (From McLean 2008)

to its focal length ($s = f_{\text{coll}}$). The field is reimaged onto the detector with a lens (or mirror) with $s' = f_{\text{cam}}$. By selecting the focal lengths of the collimator and camera lenses, one can either magnify or reduce the scale; $m = f_{\text{cam}}/f_{\text{coll}}$. Filters of arbitrary thickness can be located in this collimated beam. Moreover, the filters can be placed near the image of the primary mirror created by the collimating optics; this is called the “pupil” image. In addition, a circular aperture or stop can be placed at the pupil image to reject stray light from outside the beam. A pupil stop is important, especially in infrared cameras where the pupil is at cryogenic temperatures and so it becomes a cold stop.

Matching the spatial or spectral resolution to the physical size of the detector pixels is important. There are two factors to consider: (1) maximizing observing efficiency and (2) obtaining accurate brightness measurements (photometry). In general, the image is either critically sampled, meaning that there will be about 2 pixels (the Nyquist limit) across the resolution element, or it will be oversampled, implying that there may be about 5 pixels across the resolution element. In a spectrometer, the width of the entrance slit is usually the determining factor.

The plate scale of the telescope is given in seconds of arc per mm ($''/\text{mm}$) by

$$(\text{ps})_{\text{tel}} = \frac{206,265}{f_{\text{tel}}} \quad (12.1)$$

Here, f_{tel} is the focal length of the telescope in millimeters ($f_{\text{tel}} = D_{\text{tel}} \times F$ where F is the focal ratio or f/number) and the numerical factor is the number of seconds of arc in 1 radian. For direct imaging, the angle on the sky subtended by the detector pixel is

$$\theta = (\text{ps})_{\text{tel}} d_{\text{pix}} \quad (12.2)$$

where d_{pix} is the physical pixel size in mm; pixels are usually square. Calculating the required magnification factor can be done as follows:

- choose a value for the diameter of the seeing in seconds of arc, θ_{see}
- decide on the sampling ($p = 2\text{--}5$ pixels)
- divide seeing diameter by sampling factor to get angular size of 1 pixel, $\theta_{\text{pix}} = \theta_{\text{see}}/p$
- derive the plate scale at the detector from $(\text{ps})_{\text{det}} = \theta_{\text{pix}}/d_{\text{pix}}$
- required magnification (m) is then

$$m = \frac{(\text{ps})_{\text{tel}}}{(\text{ps})_{\text{det}}} \quad (12.3)$$

where $m = f_{\text{cam}}/f_{\text{coll}}$ as before.

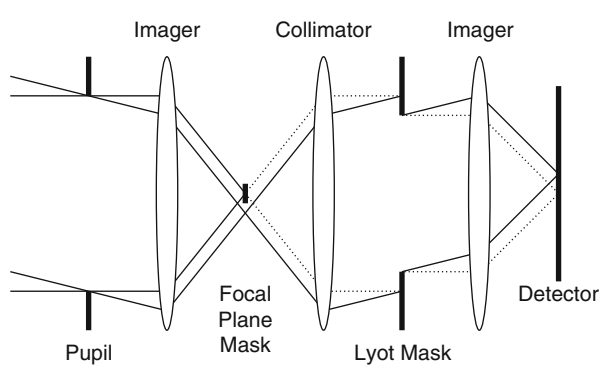
Note that m also defines an effective focal length ($\text{EFL} = mf_{\text{tel}}$) for the entire optical system. If $m > 1$, then the optical components are a magnifier, whereas if $m < 1$ (the usual case), then the optics are called a “focal reducer.” We can also relate the pixel size in seconds of arc to the f -number of the focal reducer optics (or simply, “the camera”) by

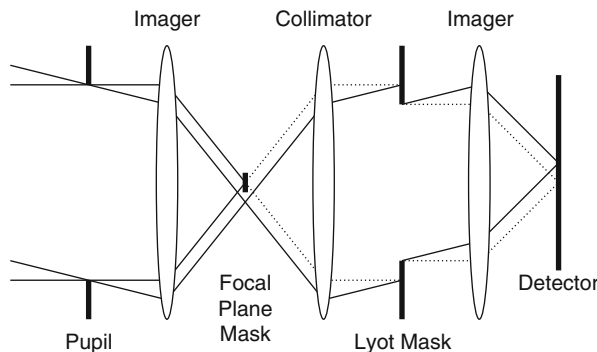
$$\theta_{\text{pix}} = 206,265 \frac{d_{\text{pix}}}{D_{\text{tel}}(f/\text{number})_{\text{cam}}} \quad (12.4)$$


where $(f/\text{number})_{\text{cam}} = f_{\text{cam}}/D_{\text{cam}} = F_{\text{cam}}$.

For example, if $d_{\text{pix}} = 18\mu\text{m}$ and $D_{\text{tel}} = 10\text{ m}$, then $\theta_{\text{pix}} = 0.37''/(f/\text{number})_{\text{cam}}$. Assuming seeing of $0.6''$ and 3-pixel sampling, this implies $\theta_{\text{pix}} = 0.2''$ which leads to $F_{\text{cam}} = 1.856$.

2.2 Coronagraphs

Astronomical investigations often seek to measure emission from material that lies at a small angular separation from a much brighter source. Distinguishing light from the much fainter source is the observational challenge. Coronagraphs attempt to suppress or steer the transmission of initially on-axis light while simultaneously allowing off-axis light to transmit relatively unimpeded. Reducing the unwanted light subsequently reduces the noise in measurements of the off-axis brightness. Many coronagraphic devices in astronomical instruments use a combination of a focal plane mask and a Lyot stop. As shown in  Fig. 12-2, a traditional Lyot coronagraph uses an opaque disk as the focal plane mask together with an undersized circular pupil as the Lyot stop or mask (Lyot 1939).



 Fig. 12-2

In the Lyot coronagraph, a focal plane mask is used to block the on-axis light, while the Lyot mask is a pupil stop that suppresses initially on-axis light that is diffracted by the pupil stop and focal plane mask. Initially off-axis light is transmitted through the system, although the intensity is reduced compared to a normal image by the action of the Lyot mask in particular

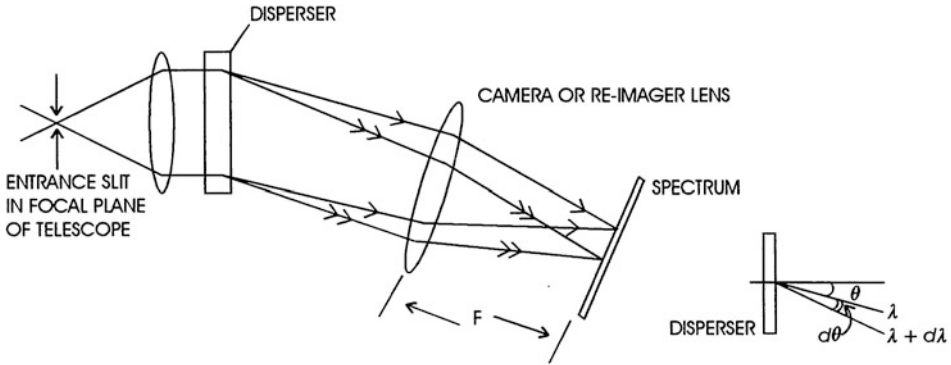
An image of a bright object, such as a star, is not an infinitesimal point in the telescope image plane. Even in the absence of wavefront errors, diffraction by the telescope pupil causes the image of a point-like source to be extended. For example, an image of a point source formed by a system with a circular pupil of diameter D will have an Airy pattern (see [Chap. 6](#)), characterized by a central core of width $\sim \lambda/D$ surrounded by concentric rings. An opaque circular focal plane mask can be used to block light from the Airy core and central bright Airy rings. By itself, however, this focal plane mask has done little to improve the ability to detect faint emission from around the star. Initially on-axis starlight is diffracted by the telescope pupil to positions beyond the extent of a finite focal plane mask, that is, to the outer Airy rings. In the pupil plane, located after the focal plane mask, this light is concentrated near the edge of the pupil image. By placing an undersized stop at this location (the Lyot mask), on-axis starlight that was initially diffracted to locations in the outer focal plane is also blocked.

Quantifying the performance of a coronagraphic imaging system is usually done in terms of the *contrast*, which measures the detectable flux relative to that of the on-axis point source. The precise definition can vary, but for point sources, this is commonly given in terms of the flux ratio between the on-axis point source *if it were not occulted* and the minimum detectable off-axis source flux. Contrast will be a function of position in the focal plane. The architecture of the coronagraph has implications for the contrast. For example, the size of the focal plane mask sets an *inner working angle*, inside of which off-axis sources are also occulted and thus undetectable. The diameter of the Lyot mask affects the amount of diffracted light from the on-axis source that is transmitted to the detector. Smaller Lyot masks result in more suppression of light from on-axis sources. However, this will not necessarily increase the achieved contrast, because decreasing the Lyot mask size also decreases the transmitted flux of any off-axis sources.

Variations of the Lyot coronagraph exist which use alternative focal plane masks. These masks manipulate the phase of the light rather than the amplitude, in order to steer it to regions in the Lyot plane that are masked. The Four Quadrant Phase Mask and the Optical Vortex Coronagraph are two examples of such architectures (Rouan et al. 2000; Palacios 2005). These devices have the advantage of reducing the inner working angle of the system. The effectiveness of diffraction control can be enhanced by adding an *apodizer* to the pupil prior to the focal plane mask. An apodizer smoothly tapers the transmission in the pupil, which has the effect of reducing the amplitude of rings in the focal plane compared to a hard-edged pupil stop (Soummer et al. 2003). Guyon (2006) reviews the performance limitations of interferometric and Lyot-style coronagraph designs. *Shaped-pupil coronagraphs* rely on novel shapes in the pupil plane to create regions in the focal plane that are devoid of diffracted on-axis light (Kasdin et al. 2003). Such systems often do not require smooth tapering of transmission profiles in the pupil mask, which can be difficult to manufacture precisely, and also do not require focal plane masking. Only a portion of the image plane achieves high contrast, which is a disadvantage.

2.3 Spectrometer Systems

[Figure 12-3](#) shows the essential features of a classical UV/O/IR spectrometer in which light enters through a narrow slit. The width of the slit must be matched to either the seeing conditions or the diffraction disk. As the beam diverges from the slit plane, it is collimated and directed to the dispersing system, after which the spectrally dispersed beam is collected by the camera optics and re-imaged onto the detector. Long-slit, multislit, and slitless spectrometers



■ Fig. 12-3

Essential features in the optical layout of a spectrometer are illustrated. The beam is collimated before intersecting the dispersive element and then the spectrum is reimaged with camera optics onto the detector (From McLean 2008)

are in wide use throughout astronomy. In almost all cases, a two-dimensional detector records the spectrum. With the advent of optical fibers to collect light from many locations in the focal plane and feed it to the entrance slit, a huge multiplex advantage can be obtained which facilitates large-scale spectroscopic surveys. Following McLean (2008), for a given wavelength (λ) range, the important design quantities are (► 12.1) the resolving power ($R = \lambda/\Delta\lambda$), (► 12.2) the slit width, (► 12.3) the diameter of the collimated beam, (► 12.4) the sampling or matching of the slit width to the detector pixels, and (► 12.5) the resulting f /number of the camera system.

Linear dispersion (L.D.) relates an interval of length (dx in mm) along the spectrum to a wavelength interval ($d\lambda$ in Ångstroms or nanometers)

$$\text{L.D.} = \frac{dx}{d\lambda} = \frac{dx}{d\theta} \frac{d\theta}{d\lambda} = F \frac{d\theta}{d\lambda} \quad (12.5)$$

Here, f_{cam} is the focal length of the spectrograph camera and $d\theta/d\lambda$ is the angular dispersion of the prism or grating device. The units are usually expressed as $\text{mm}/\text{Å}$, but a more useful form is the Reciprocal Linear Dispersion which is simply the inverse of the above expression in $\text{Å}/\text{mm}$. Expressions for angular dispersion follow from the basic equations for prisms and gratings. For a diffraction grating, the equation is

$$m\lambda = d(\sin i + \sin \theta) \cos \gamma \quad (12.6)$$

where d is the spacing of adjacent grooves or slits, i is the angle of incidence of the collimated beam, θ is the angle of the emergent diffracted beam, γ is the angle out of the normal plane of incidence (usually 0° , hence $\cos = 1$), and m is an integer called the “order” of interference. For zero order ($m = 0$), $\sin \theta = -\sin i$ or $\theta = -i$. The negative sign comes from the fact that we have chosen to call i and θ positive when on the *same* side of the normal. Whenever the rays cross over the normal, the angle of diffraction is taken to be negative. With this sign convention, the equation applies when the grating is used in transmission and when the grating is used in reflection. There is an alternative form of the equation that uses a negative sign between the terms to describe a reflection grating. In that case, the angles are positive if they are on *opposite*

sides of the normal. If the medium on either side of the grating is not a vacuum, then a more general form would be $(n_1 \sin i + n_2 \sin \theta)$. From (12.6), the angular dispersion of a grating is given by

$$\frac{d\theta}{d\lambda} = \frac{m}{d \cos \theta \cos \gamma} \quad (12.7)$$

Substituting for m/d gives

$$\frac{d\theta}{d\lambda} = \frac{\sin i + \sin \theta}{\lambda \cos \theta \cos \gamma} \quad (12.8)$$

Usually $\cos \gamma \sim 1$ and therefore angular dispersion is determined entirely by i and θ for a given λ . Many combinations of m and d yield the same A.D. provided the grating angles remain unchanged. Typical “first-order gratings” ($m \sim 1$) have 300–2,400 grooves or lines/mm; the number of lines per mm is given by $T = 1/d$. Coarse-ruled reflection gratings (large d) can achieve high angular dispersion by making i and θ very large, typically 60° . Such gratings are called “echelles” and have groove densities from 20 to 200 lines/mm with values of m in the range 10–100. This results in severe overlap of the orders unless a second disperser of lower resolving power is used at right angles to the first to “separate” the orders.

In practice, spectrometers are usually slit-width or seeing-limited. Taking $\theta_{\text{see}} = p \times \theta_{\text{pix}}$, where p is the number of pixels across the slit image, and converting to seconds of arc, gives a form which shows explicitly the trade-offs of size versus resolution:

$$R = \left(\frac{\sin i + \sin \theta}{\cos i} \right) \frac{D_{\text{coll}}}{D_{\text{tel}}} \frac{206,265}{p \theta_{\text{pix}}} \quad (12.9)$$

This formula makes it clear that as telescopes get larger, the spectrograph (defined by the beam size D_{coll}) gets larger too, all else being equal.

By tilting the facets of a reflection grating through an angle θ_B (known as the *blaze* angle) with respect to the plane of the grating surface, it is possible to maximize the grating efficiency in the direction in which light would have been reflected in the absence of diffraction. Grating efficiency is a maximum when the angle of incidence and angle of diffraction are related by $(i + \theta) = 2\theta_B$. The separation between the beams $(i - \theta)$ is just the spectrograph angle ϕ . Thus,

$$m\lambda_B = 2d \sin \theta_B \cos(\phi/2) \quad (12.10)$$

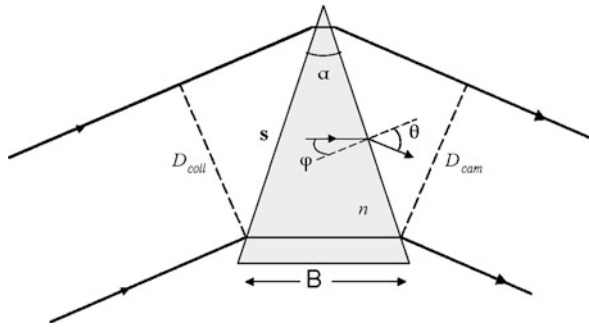
A special case occurs when $\phi = 0$, for then the incident ray enters along the normal to the facet and the diffracted ray leaves along the same direction. This is the “Littrow” condition and the incident and diffracted angles measured relative to the grating normal are now equal to each other and to the blaze angle. The grating equation simplifies to $m\lambda_B = 2d \sin \theta_B$, and the resolving power is given by

$$R = \frac{2D_{\text{coll}} \tan \theta_B}{\phi D_{\text{tel}}} \quad (12.11)$$

The only way to work in the Littrow condition is with a central obscuration in the optics. Alternatively one can use the “near” Littrow condition by moving off by a $10\text{--}20^\circ$ or the “quasi” Littrow condition by going out of the plane ($\gamma > 0^\circ$).

Prisms find applications in spectrographs both in the role of primary disperser in (usually) low-resolution instruments and as a cross-disperser in high-resolution echelle spectrographs. The basic layout of a prism disperser is shown in Fig. 12-4. From the definition of angular dispersion:

$$\frac{d\theta}{d\lambda} = \frac{d\theta}{dn} \frac{dn}{d\lambda} = \frac{B}{D_{\text{cam}}} \frac{dn}{d\lambda} \quad (12.12)$$




■ Fig. 12-4

The relationship of angles and lengths in a prism at minimum deviation are used to derive the resolving power (From McLean 2008)

In this expression, $dn/d\lambda$ describes the wavelength dependence of the refractive index n . The purely geometric term $d\theta/dn$ can be derived by differentiating Snell's law applied to the second surface and then doubling the rate to account for both surfaces. At minimum deviation, the angle $\phi = \alpha/2$ giving $d\theta/dn = [2s \sin(\alpha/2)/s \cos \theta]$. However, $2s \sin(\alpha/2) = B$, the base length of the prism, and $s \cos \theta = D_{\text{cam}}$ the emergent collimated beam width toward the camera. The resolving power of a prism is $R = B (dn/d\lambda)$, and for a slit-limited instrument, the resolving power is given by

$$R = \frac{\lambda}{\theta_{\text{res}} D_{\text{tel}}} B \frac{dn}{d\lambda} \quad (12.13)$$

A popular way to convert a camera into a spectrograph is to deposit a transmission grating on the hypotenuse face of a right-angled prism and use the deviation of the prism to bring the first order of diffraction on axis. Such a device is called a "grism" and the basic geometry (not to scale) is illustrated in  Fig. 12-5. A grism can be placed in a filter wheel and treated like another filter. The basic relationships needed to design a grism are

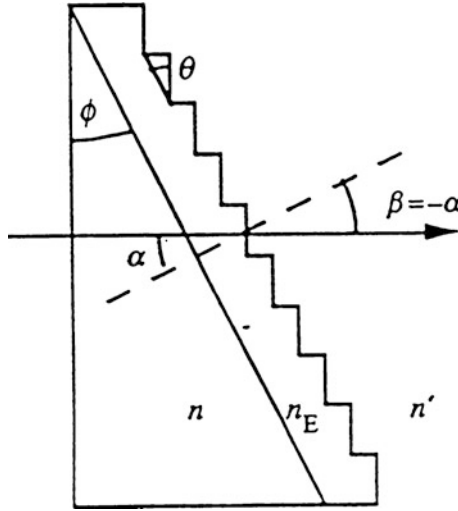
$$m\lambda_c T = (n - 1) \sin \phi \quad (12.14)$$

and

$$R = \frac{\text{EFL}}{2d_{\text{pix}}} (n - 1) \tan \phi \quad (12.15)$$

where λ_c is the central wavelength, $T (= 1/d)$ is the number of lines per mm of the grating, n is the refractive index of the prism material, and ϕ is the prism apex angle. EFL is the effective focal length of the camera system, and d_{pix} is the pixel size. The factor of 2 assumes that 2 pixels are matched to the slit width. In practice, the number of free parameters is constrained by available materials and grating rulings, and given conditions within the camera system. Resolving powers (2 pixels) of $R \sim 500\text{--}2,000$ are practical.

Most astronomical gratings are of the surface relief kind in which the grooves are formed on the surface of the substrate (direct ruling) or as a replicated grating in a material bonded to the substrate. Reflection gratings can be coated with a reflective surface such as silver or gold, where the latter is particularly useful in the infrared. An alternative technology for grating fabrication is the Volume Phase Holographic or VPH grating. Not to be confused with a normal holographic grating which is another method for creating a surface relief grating, a VPH grating



■ Fig. 12-5

A simplified schematic giving the basic geometry of a *grism* – a transmission diffraction grating deposited on the hypotenuse face of a right-angled prism (From McLean 2008)

is an optical substrate in which the refractive index varies periodically throughout the body of the grating (Barden et al. 2000; Baldry et al. 2004). The grating body is made from a thin (3–30 μm) slab of dichromated gelatine (DCG) trapped between glass plates. Light passing through a VPH transmission grating obeys the following grating equation:

$$m\lambda = n_i \Lambda_g (\sin \alpha_i + \sin \beta_i) \quad (12.16)$$

where m is an integer representing the order, n_i is the refractive index of the medium, Λ_g is the grating period (equivalent to groove spacing) and is the projected separation between the fringes; $\Lambda_g = \Lambda / \cos \varphi$ where φ is the slant angle between the grating normal and the plane of the fringes. The angles of incidence (α_i) and diffraction (β_i) are relative to the grating normal with the convention that zero order (no diffraction) corresponds to $\beta_i = -\alpha_i$. The equation applies to each layer, where $i = 0$ is the air, $i = 1$ is the glass substrate, and $i = 2$ is the DCG layer. High diffraction efficiency can occur when light is effectively reflected from the plane of the fringes, that is, when $\beta_2 + \varphi = \alpha_2 - \varphi$ in the DCG layer. This behavior is the same as Bragg diffraction of X-rays from the atomic layers in a crystal lattice. In both cases, because the thickness of the medium is much greater than the wavelength, constructive interference occurs for radiation scattered in that direction. The Bragg condition implies

$$m\lambda = 2n_2 \Lambda \sin \alpha_{2b} \quad (12.17)$$

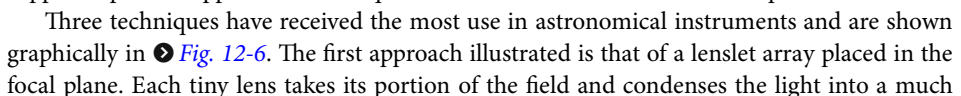
where n_2 is the refractive index of the DCG layer and α_{2b} is the “Bragg angle” or angle of incidence with respect to the plane of the fringes $\alpha_{2b} = \alpha_2 - \varphi$. At wavelengths sufficiently displaced from the Bragg condition, there is no diffraction. Diffraction efficiency also depends on the semiamplitude of the refractive index modulation (Δn_2) and the grating thickness (d). The DCG holds a fringe pattern generated by holography which provides planes of constant refractive index separated by a length $\Lambda = 1/\nu_g$. The index variations are the result of density variations

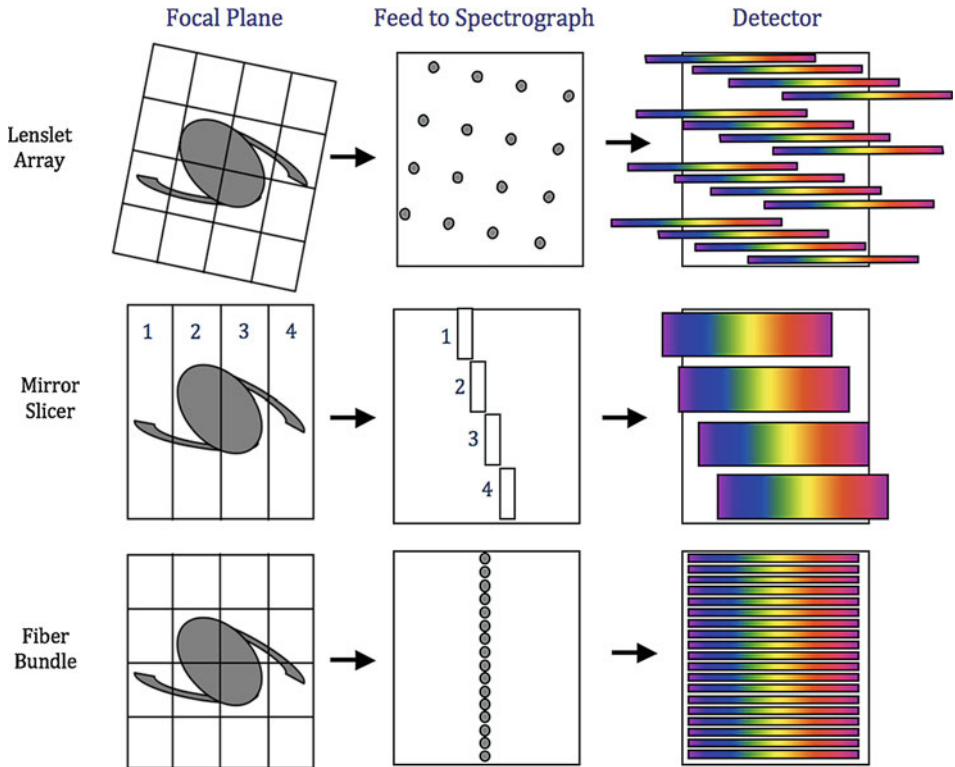
which are trapped into the material by exposure to light (the fringe pattern) because those regions collapse to a different density in the process of swelling the gelatin with water and then drying it rapidly. One form for the refractive index is $n_2(x, z) = n_2 + \Delta n_2 \cos[2\pi\nu_g(x \sin \gamma + z \cos \gamma)]$, which gives the variation in the x, z plane where z is the optical axis through the VPH, and γ is the angle between the normal to the planes and the z -axis. Line densities can range from 300 to 6,000 lines/mm, and index modulations of 0.02–0.1 are typical. Because of the Bragg condition, it is necessary to articulate the camera to a new angle to tune to a new wavelength.

2.4 Integral Field Spectrometers

An integral field spectrograph (IFS) samples a rectangular or other two-dimensional field of view and produces spectra for each spatial location. It does this by reformatting the focal plane in such a way that a traditional dispersing spectrograph can simultaneously disperse adjacent regions of the field without overlapping on the detector. Spectra from the various pieces of the field can then be reassembled into a cube of data covering two spatial dimensions and one of wavelength. Depending on the choices of field of view, spectral resolution, and spectral bandwidth, such cubes can contain tens, hundreds, or even thousands of spectral channels, all taken simultaneously and all covering a contiguous field. Almost all integral field spectrographs at telescopes use a standard CCD or infrared array as the detector. Because practical detectors are limited in terms of the total number of pixels, and since each spatial location uses hundreds or thousands of pixels for their respective spectra, the field of view of an IFS is often much smaller than in a traditional camera.

Nevertheless, having adjacent and simultaneous spectra can be a powerful scientific advantage in many cases. Examples include, measuring line ratios across such diverse objects as a high-redshift galaxy or a Jovian moon in an individual exposure in which slit losses from a classical spectrometer, changing seeing conditions and effects from adaptive optics on the point spread function are all minimized. In crowded fields like the Galactic Center, an integral field spectrometer can be used without fear of pointing errors since the entire field is reconstructed and synthetic apertures can be used on the cubes of data much like aperture photometry on individual images. In most cases, some portion of the field of view will also contain blank regions of sky, and thus atmospheric emission lines are recorded simultaneously with science photons. Depending on the design, a surprising benefit can be a very low wavefront error which is crucial for high-resolution applications. In particular, lenslet arrays and fiber bundles sample the focal plane prior to any of the spectroscopic optical elements like the collimators, cameras, and gratings. Consequently, image quality at the input to the reformatting optics remains the final image quality of the cubes. For example, in this way, the OSIRIS instrument at the Keck Observatory has a measured wavefront error below 25 nm across its entire field. In high-contrast applications, like the Gemini Planet Imager and Lyot Project, an integral field spectrograph with a lenslet array dramatically reduces noncommon path chromatic errors because the speckle pattern is recorded at the lenslet array where essentially no refractive elements, except for a window, have been in the beam path. Speckle coherence is maintained across a broad bandwidth which supports speckle suppression techniques to increase contrast with faint companions.

Three techniques have received the most use in astronomical instruments and are shown graphically in  Fig. 12-6. The first approach illustrated is that of a lenslet array placed in the focal plane. Each tiny lens takes its portion of the field and condenses the light into a much



■ Fig. 12-6

A summary of the three most common methods of creating an integral field spectrograph is illustrated. For each model, the image plane is shown on the *left*, the spectrograph input in the *middle*, and detector plane on the *right* (Based on a similar figure in Allington-Smith and Content (1998))

smaller pupil image. Thus, approximately one focal length behind the lenslet array, a grid of small pupil images serves as the entrance plane to a traditional spectrograph. The lenslet array is rotated compared to the dispersion axis of the spectrograph so that spectra are interleaved on the detector. Traditionally, lenslet-based spectrographs at optical wavelengths have been optimized for large spatial coverage with either low spectral resolution or small wavelength coverage. Lenslet designs were first proposed by Courtes (1982) and have since been implemented in a variety of instruments (e.g., Bacon et al. 1995). The first diffraction-limited use of a lenslet-based spectrograph was OSIRIS (Larkin et al. 2006) working in the near-IR with the Keck Adaptive Optics System (Wizinowich et al. 2000). This integral field spectrograph can produce spectra with as many as 1,800 spectral channels from 1,000 spatial locations in a 16×64 pattern. Similar lenslet based spectrographs with many more spatial elements and much shorter spectra ($R \sim 45$) have been chosen and built for the high-contrast Lyot Project (Hinkley et al. 2008) and Gemini Planet Imager projects. For the latter, a 200×200 lenslet yields $\sim 40,000$ spectra with each spectrum having only 18 spectral channels, but covering a full 20% bandpass.

Advantages of a lenslet-based IFS include excellent image quality because only optics in front of the lenslet array can affect the wavefront error. Lenslet arrays are also manufactured

in very large formats allowing nearly infinite expansion of the field of view. These commercial arrays also have no internal surfaces, approximately 98% fill factors and high transmission. It is generally easy to change the plate scale of a lenslet-based IFS because they are essentially like detector pixels in the sense that reimaging optics in front of them can be used to magnify the field, and their fast focal ratios make them relatively insensitive to the input focal ratio. Among the disadvantages of a lenslet-based IFS is the complexity of the data. Thousands of very long spectra are scattered over the detector in a complex pattern which must be mapped and individually calibrated. In principle, this is comparable to reductions of multislit data and has the advantage that unlike slits which can be repositioned, the lenslets have a fixed geometry. But with dense packing of spectra, sophisticated algorithms are often required to extract all of the spectra and reassemble the final cube. Lenslet array spectrographs must put at least some pixel gap between neighboring spectra. So the total number of elements in the final data cube will be at most half the number of original detector pixels. The other methods suffer similar pixel issues but for different reasons. A challenging problem with a lenslet IFS which must be considered carefully is crosstalk between neighboring spectra on the detector. Since they are staggered in wavelength space, a bright line from one spectrum could blend into a neighboring spectrum at a different wavelength creating in effect a ghost line. With good optics and knowledge of the packing geometry, this contamination can be eliminated or at least understood, but it does put pressure on the optical design, detector utilization, and data reduction algorithms.

A second technique, and perhaps the most commonly implemented one, is an image slicer. A parallel set of narrow mirrors is used to divert different parts of the focal plane into different directions. A second set of mirrors redirects these portions back into something close to a slit-like arrangement (a pseudo slit), which is then fed into a traditional spectrograph. The first cryogenic version of such an instrument for astronomy was the 3D Spectrograph (Krabbe et al. 1997). Other early successful instruments are PIFS at Palomar (Murphy et al. 1999) and SPIFFI (Tecza et al. 1998) for the VLT. The latter can be fed by an adaptive optics system where the combination is called SINFONI. For diffraction-limited spectrographs, there are some significant disadvantages to slicer-based spectrographs. The first is the relatively large size of the slicing mirrors which are usually ~ 1 mm wide. For small plate scales, this forces the design to long focal ratios which can be cumbersome. The second disadvantage is that diffraction off the individual slits causes the pupil on the grating to grow through the same process that causes pupil diffraction in lenslet-based designs. But perhaps the most important difficulty is with image quality. Particularly in the long axis of the slit, all optical elements, including the grating, affect the wavefront. However, if the wavefront error issue can be overcome, as some recent designs for SNAP and JWST indicate, then a slicer offers two very major advantages. First, it is more efficient in utilizing detector pixels (maybe by a factor of 2 or more). This translates directly into the number of spatial pixels or spectral bandwidth of the spectrograph. Secondly, each source is spread over fewer pixels which mitigates the effect of detector noise and thus increases sensitivity. In a simple comparison, the sensitivity gain could be as much as 40% at detector limited flux levels.

The third technique uses a bundle of fiber optic cables arranged to sample a region of the focal plane. This bundle is then reorganized into a linear arrangement in order to produce a slit-like pattern to feed the spectrograph. Optical fibers require a relative thick cladding and so the fill factor is usually quite low. Fibers also have fast output focal ratios making the collimator and camera optics more difficult. Fiber-based IFS systems have been used for relatively bright targets. Early examples of such spectrographs include 2D-FIS and HEXAFLEX (Garcia et al. 1994) and INTEGRAL at WHT (Arribas et al. 1998). Many of the inherent problems with fibers

can be overcome by using lenslets coupled to the input and sometimes the output of the bundles to give them higher fill factors in the focal plane. Focal ratio degradation within the fibers and coupling losses at the lenslet-fiber boundary still need to be considered.

2.5 Polarimeter Systems

To measure polarization properties, such as the fraction polarized, the direction of vibration, and the handedness of rotation, all polarimeters “convert” the polarization information into brightness modulations which are directly measurable with an electronic detector. Polarimeters can be created from camera and spectrometer designs by adding a polarization modulator. A typical approach is to construct the polarization modulator in two parts. A retardation device comes first to introduce a known and controllable phase shift into the beam, and this is followed by a fixed polarizer (also called an analyzer) that only allows one plane of polarization to pass unhindered and reduces others by the factor $\cos^2 \theta$, where θ is the angle between the polarizer’s axis and the plane of polarization in the beam. The intensity transmitted by the analyzer is therefore modulated by the action of the phase retardation device.

Linear polarization is described by three parameters: intensity (I), degree (or fraction) of linear polarization (p), and the direction of the (fixed) plane of vibration projected on the sky (θ). Circular polarization is similarly described by three parameters: intensity (I), degree of circular polarization (q), and handedness of the rotation of the electric vector (+ or –). A more convenient way to express polarization information is to use the four Stokes parameters (I , Q , U , V). These quantities are phenomenological, that is, they are more directly related to actual measurements. The Stokes parameters are easily related to the amplitudes (E_x , E_y) of the electric vector in two orthogonal directions and to the phase difference (δ) between the two components (e.g., Clarke and Grainger 1971; Tinbergen 1996). The degree of linear and circular polarization is given by

$$p = \frac{[Q^2 + U^2]^{\frac{1}{2}}}{I}, q = \pm \frac{V}{I} \quad (12.18)$$

and the direction of vibration of the linearly polarized part is given by

$$\tan 2\theta = \frac{U}{Q} \quad (12.19)$$

and it follows that

$$\begin{aligned} Q &= Ip \cos 2\theta \\ U &= Ip \sin 2\theta \\ V &= Iq \end{aligned} \quad (12.20)$$

The intensity of light transmitted by a retarder of retardance τ at angle ψ followed by a perfect polarizer with principal plane at $\phi = 0^\circ$ or $\phi = 90^\circ$ (upper/lower signs, respectively) is given by

$$I' = \frac{1}{2} [I \pm Q(G + H \cos 4\psi) + \pm UH \sin 4\psi \pm V \sin \tau \sin 2\psi] \quad (12.21)$$

where

$$G = \frac{1}{2}(1 + \cos \tau), \quad H = \frac{1}{2}(1 - \cos \tau), \quad \tau = \frac{2\pi}{\lambda} \delta \quad (12.22)$$

There are several special cases and multiple ways to solve these equations for the Stokes parameters. Although harder to measure than other properties, polarization from astronomical

sources can be detected from X-rays to radio waves. Because it may contain information about the formation of the early universe, one important area of interest is the polarization of the cosmic microwave background.

2.6 Interferometers

Interferometer techniques in astronomy are applied in two different ways, one as a collection method and the other as a detection method. Combining the light collected by many widely separated telescopes overcomes the diffraction limit of an individual telescope. Single-aperture telescopes can be equipped with interferometer equipment for specific detection purposes. Several types of detection interferometers have been used for spectroscopy, such as the Fourier transform spectrometer (FTS) which is a scanning Michelson interferometer and the Fabry-Perot interferometer which is an imaging spectrometer.

A typical FTS is shown in **Fig. 12-7**. For a collimated monochromatic beam, the intensity at the detector is determined by the “path difference” $\Delta x = 2(x_b - x_a)$, where x_a refers to the arm containing the fixed mirror A and x_b is the distance to the scanning mirror B. The phase difference is given by $k\Delta x$ where $k = 2\pi/\lambda$. The fraction of the incident beam in the output is given by

$$T(k, \Delta x) = \frac{1}{2}[1 + \cos(2k\Delta x)] \quad (12.23)$$

from which it follows that $T = 1$ when the combining beams are in phase and $T = 0$ when they are 180° out of phase. Given an incident beam whose spectrum is $I(k)$, the signal F measured in the output is

$$F(\Delta x) = c \int I(k) T(k, \Delta x) dk = \text{constant} + \frac{c}{2} \int I(k) \cos(2k\Delta x) dk \quad (12.24)$$

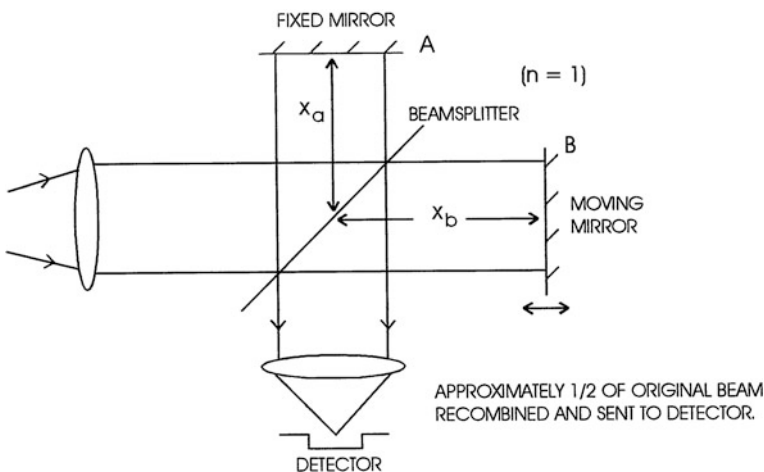


Fig. 12-7

The principle of the scanning Michelson interferometer is shown. As the mirror is scanned, the intensity recorded by the detector is modulated to produce an interferogram. The spectrum can be extracted by an inverse Fourier transform (From McLean 2008)

where c is a constant. The measured signal $F(\Delta x)$ is called the interferogram, and the last integral is the Fourier cosine transform of the spectrum. Therefore, the transform of the interferogram is $I(k)$.

The Fabry-Perot interferometer is an imaging spectrometer formed by placing a device called an “etalon” in the collimated beam of a typical camera system. One arrangement is shown in **Fig. 12-8**. The etalon consists of two plane parallel plates with thin, highly reflective coatings on their inner faces. The plates are in near contact but separated by a distance d . Assuming that the refractive index of the medium in the gap is n (usually $n = 1$) and θ is the angle of incidence of a ray on the etalon (usually very small), then multiple reflections and destructive interference within the gap occur and the wavelengths transmitted with maximum intensity obey the relation

$$m\lambda = 2nd \cos \theta \quad (12.25)$$

For monochromatic light, the image is a set of concentric rings. To ensure that a sufficiently narrow band of light passes through the system, it is necessary to “prefilter” the light. This can be done with a very narrow band interference filter. Usually, a circular aperture isolates the central order which has an angular diameter $\delta\beta = \sqrt{(8/R)}$ and the free spectral range is given by

$$\Delta\lambda_{\text{FSP}} = \frac{\lambda}{m} = \frac{\lambda^2}{2nd} \quad (12.26)$$

The resolving power ($R = \lambda/\delta\lambda$) is

$$R = \frac{2Fnd}{\lambda} \quad (12.27)$$

where $F (= \Delta\lambda_{\text{FSP}}/\delta\lambda)$ is called the “finesse” of the etalon, which is a measure of the plate quality and the reflectance (r) of the coatings; $F = \pi\sqrt{r}/(1-r)$ and typical values are 30–50. Defining $\delta = (2\pi/\lambda)(2nd \cos \theta)$, the transmitted intensity is $I(\delta) = I(0)/[1 + (2F/\pi)^2 \sin^2(\delta/2)]$. One application of Fabry-Perot etalons is the Taurus Tunable Filter (Bland-Hawthorn and Kedziora-Chudczek 2003) which allows wide-field narrow-band imaging with a CCD.

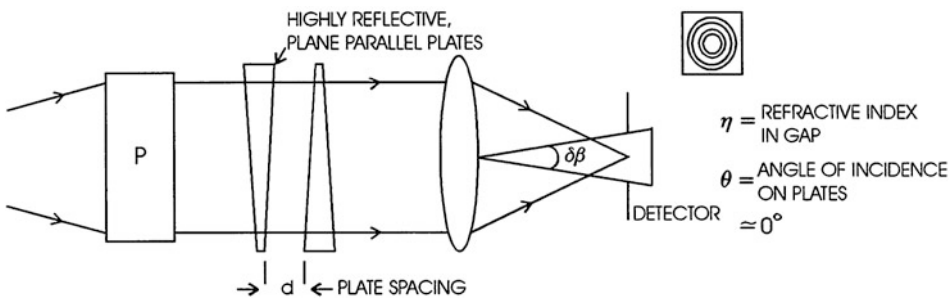


Fig. 12-8

One typical layout for a Fabry-Perot interferometer is shown. The device P is used to narrow the range of wavelengths fed to the etalon (From McLean 2008)

3 Detectors and Materials

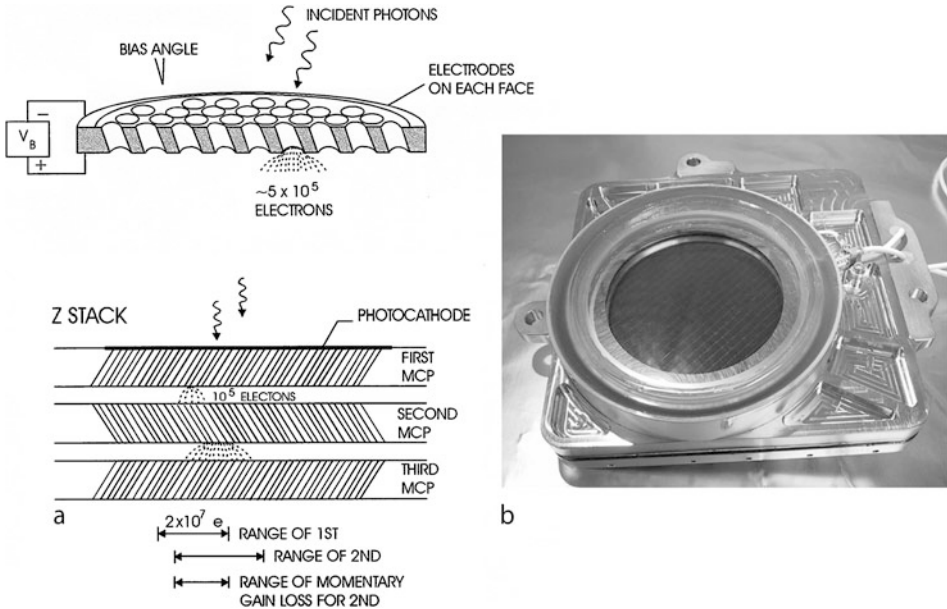
3.1 Classification of Detectors

High-energy photons are usually detected with particle detectors, but for lower energy photons, detectors are generally grouped into three broad classes:

1. *Photon detectors* in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma rays to the far-infrared.
2. *Thermal detectors* in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical conductivity; thermal detectors have a broad spectral response but are often used for infrared and submillimeter detection.
3. *Coherent detectors* in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

Photon detectors can be subdivided into (◆ 12.1) photoemission devices employing the external photoelectric effect in which the photon causes a charge carrier (electron) to be ejected from the material and (◆ 12.2) photoabsorption devices that use the internal photoelectric effect in a semiconductor to free a charge carrier within the material. The most well-known detector in the *photoemission* category is the photocathode of a photomultiplier tube (PMT) in which an electron is emitted from the photocathode surface and subsequently amplified by a cascade of impacts with secondary surfaces before being detected as a charge pulse.

Photoemissive materials can provide excellent detectors far into the ultraviolet. Most importantly, it is possible to create ultraviolet imaging devices based on this process. For example, long, narrow curved tubes or “microchannels” of lead oxide (PbO) can perform the same function as the secondary surfaces in a PMT resulting in a large pulse of electrons emerging from the end provided there is a potential gradient. As shown in (◆ Fig. 12-9a), such channels can be packaged very close together (like straws in a box) to make a two-dimensional array called a microchannel plate (MCP). MCPs are used across most of the UV and have been the detector of choice for almost all major UV missions. For example, the GALEX spacecraft employs two MCPs (◆ Fig. 12-9b) fabricated by the experimental astrophysics group at the University of California Berkeley, Space Sciences Laboratory. This group has been responsible for the development of most of the UV detectors in space. For example, there were seven on EUVE, four in two different instruments on SOHO, two in FUSE and a custom-designed MCP for COS (Hubble), to give just a partial list (Siegmond et al. 2007). The length of the microchannel is typically 50–100 times the diameter of the channel, which implies a large surface-to-volume ratio and the tendency to trap residual gas unless exceptional measures on cleanliness and plate conditioning are employed. Because MCPs are operated at potentials of a few thousand volts, residual gases can lead to destructive discharges. The channels have diameters ranging from 5 to 25 μm on 10 to 40 μm centers, and plates with active areas as large as $100 \times 100 \text{ mm}^2$ are available; the GALEX detectors are 75 mm in diameter with an active area of 68 mm. The response of the MCP is a strong function of the angle of incidence of the photons. Photocathodes can be placed on the top face or on a window in proximity focus immediately above the MCP. Materials with



■ Fig. 12-9

(a) The structure of a microchannel plate (MCP) device. (b) The GALEX MCP detector (Credit: Experimental Astrophysics Group, UC Berkeley (From McLean 2008))

large work functions such as CsI, CsTe, and KBr have good UV quantum efficiency but very low response to visible photons. More recently, gallium nitride (GaN), which has a band gap of 3.4 eV, has been added to the list of photocathodes available for UV astronomy. Microchannel plate detectors use a variety of anode structures. One of the simplest is a single resistive anode in which the location of the event is determined by the amount of charge or current “divided” between amplifiers attached to the corners. Other anode structures include the wedge and strip anode, the spiral anode, and the delay Line, each of which is described as “continuous” anodes. It is also possible to utilize “discrete” anode structures at the expense of many more amplifiers and encode the event location through direct detection. One such system is called the Multi-Anode Microchannel Array (MAMA). MAMA detectors with $1,000 \times 1,000$ pixels were constructed for Hubble’s STIS and ACS instruments.

Detectors employing *photoabsorption* make up the largest category. There are many possible outcomes, including chemical change as in photography, but absorption in semiconductor devices is the important one for astronomy. There are essentially two basic types of interactions, the photoconduction effect and the photovoltaic (or photodiode) effect. The photoconductor is composed of a single uniform semiconductor material in which the conductance is changed by the creation of free charge carriers in the material when photons are absorbed. There is usually always an external applied electric field. In the photodiode (or photojunction), internal electric fields and potential barriers are created by suitable junctions between different materials or deliberate variations in the electrical properties of the material so that photogenerated carriers in these regions respond to those fields.

3.2 Semiconductors

When individual atoms come close together to form a solid crystal, electrons in the outermost orbits or upper energy levels of adjacent atoms interact to bind the atoms together. Because of the very strong interaction between these outer or “valence” electrons, the upper energy levels are drastically altered, with the result that the outer electrons are shared between the different atomic nuclei. In fact, the energy levels are spread out into a “band.” The lowest band of energies, corresponding to all the innermost orbits of the electrons, is filled with electrons because there is one electron for each atom. This band of filled energy levels is called the “valence band.” Conversely, the upper energy band is empty of electrons because it is composed of the combined unoccupied higher energy levels of the individual atoms in the crystal. It is called the “conduction band” for reasons that will become apparent. Thus, the individual atoms have a gap in energy between the inner filled levels and the outer unoccupied levels. The energy region between the valence band and the conduction band in the crystal must be a “forbidden energy gap” (E_G). The crystal must be pure and contain atoms of only one kind; otherwise, additional energy levels corresponding to those atoms will be formed. More importantly, the periodic or repetitive crystalline structure must be unbroken to avoid distortions in the energy levels caused by abnormal sharing of electrons. In practice, both of these conditions are violated in real crystals, and such departures from the simplified model contribute to degraded performance.

In metals, the valence and conduction bands overlap, and so any of the many valence electrons are free to roam throughout the solid to conduct electricity and heat and to move in response to the force of an electric field. Insulating materials, on the other hand, have a highly ordered structure and a very wide forbidden energy gap. The conduction band is totally empty of electrons and so cannot contribute to an electrical current flow. Electrons in the completely filled valence band cannot move in response to an electric field because every nearby orbit is occupied. In a semiconductor, a few electrons can be elevated from the valence band to the conduction band across the forbidden gap merely by absorbing heat energy from the random, microscopic, jostling motions of the crystal structure at normal “room” temperature. Thermal energy is given approximately by

$$E_{\text{th}}(\text{eV}) = kT = 0.026(T/300)\text{eV} \quad (12.28)$$

where k is Boltzmann’s constant and T is the absolute temperature. At room temperature ($T = 300 \text{ K}$), the thermal energy is quite small at 0.026 electron volts. Electrons promoted to the conduction band can then conduct electricity, that is, they are free to move under the influence of an electric force field. Interestingly, the corresponding vacancies or “holes” left in the valence band allow it to contribute to electrical conductivity as well because there is now somewhere for electrons in adjacent atoms to go; descriptions of solid-state devices therefore refer to “electron-hole” pairs.

Most semiconductor crystals have band gap energies around 1 eV, but the range is from almost 0 to about 3.5 eV. Visible light photons have energies around 2.25 eV (for 550 nm). As the number of electrons which can be promoted to the conduction band by absorbing heat will vary with the temperature of the crystal, typically as $\exp(-E_G/2kT)$, those semiconductors with larger band gaps are preferred because transistors and other devices made from them will be less sensitive to environmental changes. If the semiconductor is cooled to a low enough temperature, random elevation of valence electrons to the conduction band can be virtually eliminated. The primary semiconductors, like silicon and germanium, belong to the “fourth

column” elements of the periodic table, which also includes carbon. Each of these elements has four valence electrons. Compounds of elements on either side of the fourth column can be formed, and these alloys will also have semiconductor properties. For example, gallium arsenide (GaAs) and indium antimonide (InSb) are III–V (or “three–five”) compounds, and mercury-cadmium telluride (HgCdTe) is one possible II–VI (or two–six) compound.


When a photon is absorbed in the crystalline structure of silicon, its energy is transferred to a negatively charged electron, the photoelectron, which is then displaced from its normal location in the valence band into the conduction band. When the electron reaches the conduction band, it can migrate through the crystal. Migration can be stimulated and controlled by applying an electric field to the silicon crystal by means of small metal plates called “electrodes” or “gates” connected to a voltage source. For each semiconductor, there is a wavelength of light beyond which the material is insensitive to light because the photons are not energetic enough to overcome the forbidden energy gap (E_G) in the crystal. The cutoff wavelength is given by

$$\lambda_c = \frac{hc}{E_G} \quad (12.29)$$

where h is Planck’s constant and c is the speed of light; $hc = 1.24$ for wavelengths in microns and energy in electron volts.

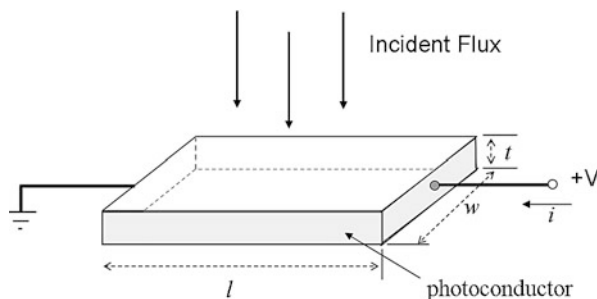
All of the materials mentioned so far are “intrinsic” semiconductors because each has a well-defined band gap intrinsic to the material. It is also possible to create an “extrinsic” semiconductor in which impurity atoms produce intermediate energy levels within the forbidden gap. For example, when silicon atoms in the crystal structure are deliberately replaced with other atoms, the semiconductor is said to be “doped.” If the impurity atom has more valence electrons than the semiconductor, then it will donate these *negative* charges to the conduction band; such a material is called *n-type*. Conversely, if the impurity atom has fewer valence electrons than the semiconductor, then a *positively* charged hole is left in the valence band ready to accept any available electrons; this material is called *p-type*. In *p-type* material, there is an excess of holes and so electrons are said to be the minority carriers of charge, whereas the opposite is true for *n-type* material. Because of the much lower transition energies, extrinsic semiconductors are used in far-infrared photon detection.

3.3 Photoconductors

This is the simplest application of a semiconductor for detection of photons. Photons are absorbed and create electron-hole pairs. If the material is extrinsic rather than intrinsic, then E_i must be substituted for E_G . Also, for extrinsic materials, there are limits on solubility of the dopants and high concentrations introduce unwanted conductivity modes. In practice, both electrons and holes contribute to the photocurrent, but it is usually the electrons that dominate. The main parameters in the construction and operation of a photoconductor are shown in  Fig. 12-10. For this discussion, it is assumed that the detector has been cooled to eliminate thermally generated charges.

The average photocurrent (I) between the terminals that is generated by an incident flux with power P (watts) is given by

$$I = (e\eta P/h\nu)(v\tau/l) \quad (12.30)$$



■ Fig. 12-10

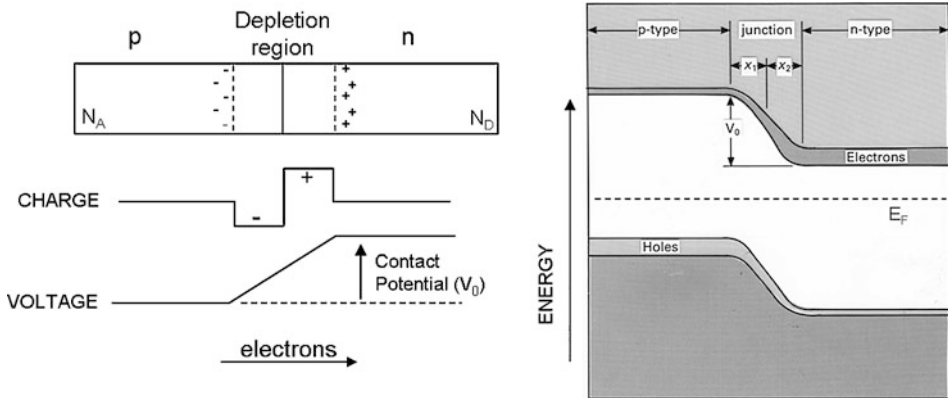
Construction and operation of a semiconductor used in photoconduction mode (From McLean 2008)

In this expression, η is the quantum efficiency and $P/h\nu$ is just the photon arrival rate. The quantity τ is called the mean carrier lifetime and measures how long the photogenerated charge exists before recombination. Values are usually less than to much less than a few milliseconds but depend on doping and temperature. The average charge carrier velocity is v , which is related to the applied electric field across the photoconductor $E = V/l$ by $v = \mu E$ where μ is called the mobility of the charge carrier. Thus, l/v is the transit time across the device from one terminal to the other, and the quantity $G = v\tau/l$ is just the ratio of mean carrier lifetime to transit time. It is known as the “photoconductive gain.” The response (S) of the detector (in amps per watt or volts per watt) is just I/P or V/RP where V is the bias voltage across the photoconductor and the resistance R due to the photocurrent is $l/\sigma A$ and the conductivity $\sigma = ne\mu$, where n is the average density of carriers. It follows that $S = (e\eta G/hc)\lambda$. Finally, the root mean square noise for a photoconductor is given by $\sqrt{(4eGIB)}$ where B is the electrical bandwidth of the measurement.

3.4 Photodiodes

Junctions between p- and n-type regions are used many times in semiconductor structures to produce different devices. One such device is the photodiode. When a pn junction is formed as shown in [Fig. 12-11](#), electrons from the n region tend to diffuse into the p region near the junction and fill up some of the positively ionized states or holes in the valence band, thus making that p-type region more negative than it was. Similarly, the diffusion of holes from the p to the n side leads to an increasingly more positive electrical potential.

A narrow region forms on either side of the junction in which the majority charge carriers are “depleted” relative to their concentrations well away from the junction. As the concentration of electrons in the n-type material is usually very much larger than in the p-type material, the flow of electrons would tend to be one way were it not for the fact that the diffusion process itself begins to build up an electrostatic potential barrier which restrains the flow of electrons from the n-type region; the buildup of electrons on the p side makes it negatively charged which starts to repel further diffusion. The magnitude of this potential barrier (V_0) depends on the impurity concentrations, that is, on the number of donor electrons at the junction that are available for transfer to nearby acceptor levels and is just equal to the required shift of the energy bands



■ Fig. 12-11

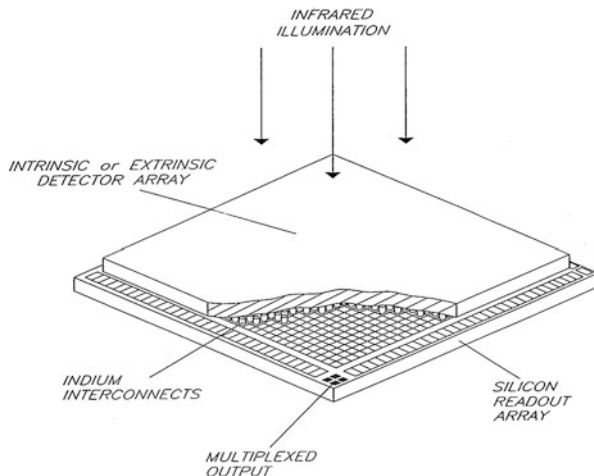
Formation of a pn junction between p-doped and n-doped materials results in a region depleted of carriers and the creation of a potential barrier (From McLean 2008)

needed to ensure that the Fermi level (E_F) remains constant throughout the crystal. The Fermi level is the energy at which there is a 50/50 chance of the corresponding electron energy state or orbit being occupied by an electron. For an intrinsic semiconductor, E_F lies halfway between the valence and conduction bands, whereas for an n-type doped semiconductor, the Fermi level moves up toward the conduction band and conversely p-type doping lowers the Fermi level.

When a positive voltage is applied to the p side of the junction, it will tend to counteract or reduce the built-in potential barrier and attract more electrons across the junction, whereas a negative voltage on the p side will enhance the internal barrier and increase the width of the depletion region; these conditions are called “forward” and “reversed” bias, respectively. Therefore, on one side of a pn junction, there is a region which is more negative, and on the other side, there is a region which is more positive than elsewhere in the crystal. When light of the correct wavelength is absorbed near the junction, an electron-hole pair is created and the potential difference across the junction sweeps the pair apart before they can recombine. Electrons are drawn toward the region of greatest positive potential buried in the n-type layer which therefore behaves like a charge storage capacitor. Of course, as more electrons accumulate, the positive potential is progressively weakened. In the photodiode, an electron-hole pair is created within the depletion region by the absorption of a photon, and the charge carriers are immediately separated by the electric field across the junction. The current due to an incident photon flux (signal and background) of power P is just $I = e\eta P/h\nu$, and the root mean square noise is given by $\sqrt{(2eIB)}$ where B is the electrical frequency bandwidth of the measurement. Comparing these results to the photoconductor shows that $G = 1$ for the photodiode and the noise is less by a factor of $\sqrt{2}$ because recombination does not occur in the depletion region.

3.5 Applications to CCDs and IR Arrays

Most infrared arrays are either photodiodes or photoconductors. Infrared arrays employ a hybrid construction. They are like a sandwich (► Fig. 12-12) in which the upper slab is the IR sensor (e.g., InSb, HgCdTe; Si:As, Ge:Ga), and the lower slab is a silicon readout device.



■ Fig. 12-12

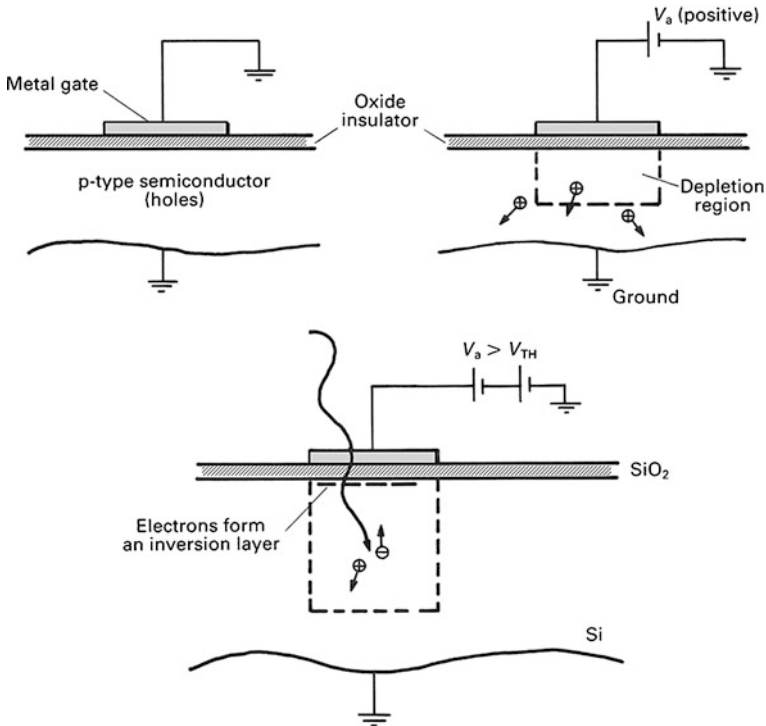
The “hybrid” structure of infrared array devices: the two slabs are separated by a grid of tiny indium bumps that remain soft at cryogenic temperatures (From McLean 2008)

The infrared layer is a tightly packed grid of individual pixels with minimum dead space between them. Both slabs are provided with a grid of electrical connections in the form of tiny raised sections – referred to as “bumps” – of an electrical conductor called indium; indium remains soft at low temperatures. The two slabs are literally pressed together to enable the indium bumps to mate. A microscopic array of “switches” made from Metal–Oxide–Semiconductor Field-Effect Transistors (MOSFETs) is used to access the signal from each IR pixel (whether photodiode or photoconductor). Charge storage may occur on the junction capacitance of the IR sensor itself (in the case of a photodiode) or on a separate storage capacitor associated with the silicon circuitry.

The entire structure is often called a Focal Plane Array (FPA) or a Sensor Chip Assembly (SCA), and the silicon readout integrated circuit part by itself is called a ROIC or mux. In the infrared array, stored charge is read out *directly* from each pixel in turn using a source follower transistor which permits nondestructive sampling of the signal voltage. The charge does not “pass through” any other pixels in the array. This is quite different from the CCD discussed below.

Silicon CCDs (charge-coupled devices) are used widely in astronomy from X-ray to near-infrared wavelengths. The CCD is an array of individual pixels each one of which can absorb photons and utilize the energy to release an electron within the semiconductor. To confine the electron within a pixel requires a special electrostatic field to attract the charged electron to a specific spot. Creating a storage region capable of holding many charges is more complicated. It is achieved by applying metal electrodes to the semiconductor silicon together with a thin (100 nm = 0.1 μm) separation layer made from silicon dioxide, which is an electrical insulator. The resulting structure behaves like a parallel plate capacitor which can therefore store electrical charge. It is called an MOS (metal-oxide-semiconductor) structure.

An electric field is generated inside the silicon slab by the voltage applied to the metal electrode. If the material is p-type, then a positive voltage on the metal gate will repel the holes which



■ Fig. 12-13

Development of a single metal-oxide-semiconductor (MOS) storage well, the basic element in a CCD, is shown for different applied gate voltages (From McLean 2008)

are in the majority and sweep out a region depleted of charge carriers. These conditions are illustrated in Fig. 12-13. When a photon is absorbed in this region, it produces an electron-hole pair, but the hole is driven out of the depletion region and the electron is attracted toward the positively charged electrode. The MOS capacitor is the combination of two parallel plate capacitors, namely, the oxide capacitor and the silicon depletion region capacitor, and therefore, the capacitance is proportional to the area of the plates (electrodes) and inversely proportional to their separation. As the voltage on the plate can be controlled, then the depletion width can be increased or decreased, and so the capacity to store charge can also be controlled. The depletion region is an electrostatic “potential well” into which many photogenerated charges can be collected. Typically, the number of electrons stored is just $Q = CV/e$, where e is the charge on the electron (1.6×10^{-19} C), V is the effective voltage, and the capacitance C is given by the “parallel-plate” formula $C = A\kappa\epsilon_0/d$ in which A is the area of the pixel or gate electrode, d is the thickness of the region, κ is the dielectric constant of the SiO₂ insulator (~ 3.9), and ϵ_0 is the permittivity of free space (8.85×10^{-12} farad/m). As the voltage on the electrode increases, the “depth” of the well increases; other ways are needed to create sidewalls to the well. Eventually, at a certain “threshold” voltage, even the minority charge carriers due to impurities, electrons from a p-type semiconductor, will be drawn to the electrode.

The unique feature of the CCD that gives it its name is the way in which the photogenerated charge, and hence the image of the scene, is extracted from the MOS storage and detection site. It is called “charge coupling.” To transfer charge from under one electrode to the area below an adjacent electrode, raise the voltage on the adjacent electrode to the same value as the first one. This is like lowering the floor of the adjacent potential well. The charges will now flow, like water, and be shared between both regions. Transfer can be in either direction, and by connecting sets of electrodes together, the entire charge stored on the two-dimensional imaging area can be moved simultaneously in that direction. When the voltage on the original electrode is reduced to zero volts, transfer is complete because the collapse of the storage well pushes any remaining charges across to the new electrode. Because it takes three electrodes to define 1 pixel, three of the above transfers are required to move the two-dimensional charge pattern by one whole pixel step along the direction at right angles to the electrode strips. The process of raising and lowering the voltage can be repeated over and over. These drive or “clock” pulses can be described in a diagram called a timing waveform. Finally, there is another set of electrodes at right angles to the first to enable charges to be moved along that row to the output amplifier where the charge is converted to a voltage that can be further amplified and digitized by an analog-to-digital converter. Details and applications can be found in Janesick (2001) and McLean (2008).

The digital signals recorded by the detector system – usually called Data Numbers (or DN) or sometimes Analog-to-Digital Units (ADU) – must be turned back into microvolts and then into electrons and finally to photons in order to calibrate the system. The relation between DN and microvolts at the CCD or infrared array output depends on the “gain” of the amplifiers in the system, and conversion between microvolts at the output and an equivalent charge in electrons requires knowledge of the capacitance (C) of the output node of the on-chip amplifier. Counts or DN recorded in a given time by the camera system are linearly related to the numbers of electrons in the charge packets by

$$S = \frac{(N_e + N_D)}{g} + b \quad (12.31)$$

where S is the recorded output signal in Data Numbers (or counts), N_e is the number of electrons in the charge packet (ηN_p), and the system photon transfer gain factor is g electrons/DN; b is the electronic offset or bias level (in DN) for an empty charge packet and N_d is the residual dark current signal still present after cooling the device. Both the bias (b) and the dark current (N_d) can be determined from measurements without illumination and can therefore be subtracted. There are two ways to derive the transfer factor g in electrons/DN, either by calculation, knowing the overall amplifier gain and the capacitance of the CCD/IR array, or by a series of observations of a uniformly illuminated scene at different brightness levels.

Let V_{fs} be the full scale voltage swing allowed on the A/D unit, and n be the number of bits to which the A/D can digitize. The full scale range is therefore subdivided into 2^n parts, the smallest part – the least significant bit or LSB – is simply 1 DN. Thus, the voltage corresponding to 1 DN at the A/D unit is $V_{fs}/2^n$; as an example, suppose the full scale voltage is 10 V and the A/D is 16 bits then 2^n is 65,536 and so the ratio is 0.0001525 V, or 152.5 μ V at the A/D is equivalent to 1 DN. Similarly, for a 14 bit A/D, the range is 16,384 and 1 DN corresponds to 610 μ V. To get the number of microvolts corresponding to 1 DN at the CCD rather than at the A/D, divide the number derived above by the total gain product A_g of all the amplifiers in the system; usually this means the on-chip amplifier (A_{SF}), the preamplifier (A_{pre}), and a postamplifier (A_{post}). To convert this number of microvolts to an equivalent charge of electrons, multiply by the CCD

capacitance (C) and divide by the value of the charge on the electron (e). Therefore,

$$g = \frac{V_{fs}C}{2^n A_g e} \quad (12.32)$$

where $e = 1.6 \times 10^{-19}$ Coulombs.

Assuming that there are only two sources of noise when imaging a uniformly illuminated flat field, photon noise from the signal and readout noise from the detector, these two noise sources should be independent and random. Therefore, adding them together in quadrature gives the total noise

$$(\text{noise})^2 = p^2 + R^2 \quad (12.33)$$

This expression applies to photoelectrons and not to counts (DN). The measured quantities, the mean signal (S_M) and its variance (V_M), are in DN. To convert from electrons to DN in (► 12.33), divide each noise term by g (electrons/DN) to give

$$\left(\frac{\text{noise}}{g}\right)^2 = \left(\frac{p}{g}\right)^2 + \left(\frac{R}{g}\right)^2 \quad (12.34)$$

The left-hand side is now exactly V_M , the observed variance in DN. Also, the mean number of photoelectrons is $g(e^-/\text{DN})S_M(\text{DN})$ or gS_M , and the photoelectron noise (p) on this number is simply the $\sqrt{(gS_M)}$ for Poisson statistics, so $p^2 = gS_M$. Hence, (► 12.34) becomes

$$V_M = \frac{1}{g}S_M + \left(\frac{R}{g}\right)^2 \quad (12.35)$$

► Equation 12.35 is just the equation of a straight line in a signal-variance plot of $y = V_M$ and $x = S_M$. Plotting these observed quantities (noise-squared and signal) as the illumination changes will yield a straight-line of gradient (slope) $m = 1/g$ with the value of the intercept on the V_M axis when $S_M = 0$ giving $(R/g)^2$ which yields R as g is known from the slope.

3.6 Detectors for High Energy

Gamma rays cannot be readily focused, even using the grazing incidence mirrors that operate successfully for X-rays. In addition, gamma rays are different from other photons in that they may not be fully absorbed. In fact, there are three regimes: total absorption by the photoelectric effect, Compton scattering, and pair production. Pair production dominates above 10 MeV. Pair conversion telescopes use devices like spark chambers and silicon strip detectors to track the particles produced. Sheets of silicon detectors can be stacked in towers to produce the solid-state equivalent of a spark chamber. An incident gamma-ray photon is forced to pair-produce by a plate of high atomic weight material, and the electrons and positrons from the conversion cause ionization in the silicon strip. Another useful device is the proportional counter. Cylinders of solid germanium with a central axial node and surrounded by a cylindrical cathode. The incoming gamma ray creates ion-electron pairs, and the electrons are attracted to the anode such that the number released is proportional to the gamma ray's energy. These kinds of units can form a "pixel" in a large array of germanium detectors. Another semiconductor material with good stopping power is cadmium zinc telluride, sometimes called simply CZT. When ionizing radiation interacts with the CZT crystal, electron-hole pairs are created in proportion to the energy of the incoming photon. CZT elements can be tightly packed to make an

array detector for gamma rays. A large array of CZT elements is used, for example, in the Burst Alert Telescope (BAT) on the *Swift* satellite (Gehrels et al. 2004). Finally, Earth's atmosphere forces high-energy gamma rays to pair-convert, and the resulting shower of particles produce Cherenkov light which can be detected by wide-field telescopes equipped with "cameras" using hundreds of PMTs as the pixel elements.

3.7 Thermal Detectors

In the class of thermal detectors, there is only one type that is used widely in astronomy and that is the "bolometer," which is described in more detail in [Chap. 14](#). Semiconductor bolometers, based on silicon or germanium, are well developed for far-infrared and submillimeter astronomy. Essentially, a bolometer consists of a sensitive thermometer and a high cross-section absorber that absorbs almost all of the incident radiation falling on it, that is, QE (η) \sim 100%. The absorber has a heat capacity of C joules per kelvin. The thermometer and absorber are connected by a weak thermal link to a heat sink at a low temperature, and the conductivity of this link is G watts per kelvin. If the detector element of the bolometer absorbs an amount of energy $E = \eta P \Delta t / h\nu$ in a time interval Δt from a source with power P, that energy is converted to heat which raises the temperature by an amount $\Delta T = T - T_0 = E/C$ above that of a heat sink at T_0 . The temperature rise decays exponentially as power in the absorber flows out to the heat sink via the weak link and the time constant is $\tau = C/G$. Temperature rise is proportional to the absorbed energy. In the classical circuit, a constant bias current, generated by the bias supply and a load resistor, flows through the bolometer. Provided that the bias power remains constant

$$T = T_0 + (P + P_{\text{bias}})/G \quad (12.36)$$

and the temperature rise causes a change in bolometer resistance, and consequently a change in the voltage across it which can be measured. Differential on/off source measurements are normally required to remove bias levels. A more detailed account is given in Rieke (2003). Arrays of bolometers now exist and will be described in the chapter on far-infrared detectors.

3.8 Coherent Detectors

A coherent detector or "receiver" is any device that directly responds to the electric field of the wave. The most important form of receiver is the heterodyne or superheterodyne which functions by mixing signals of different frequencies and detecting a signal at the difference or "beat" frequency between the original two frequencies. Depending on the frequency of the incoming wave, the electric field may be converted to an electrical signal which is then amplified before being mixed with a local oscillator. For frequencies below 1 GHz, cryogenic transistor preamplifiers are used. From 1 to 40 GHz, FET, parametric and maser amplifiers are employed, but above 40 GHz, the mixer must precede the preamplifier in order to reduce the frequency before amplification. The local oscillator (LO) produces a strong signal with a frequency that is close to but different from the signal frequency. The beat frequency, or intermediate frequency (IF), is $\nu_{\text{IF}} = \nu_{\text{S}} - \nu_{\text{LO}}$. Subsequent amplification and filtering of the intermediate frequency signal by a large factor then follows and the resultant signal is rectified by a diode and integrated. The key element in this entire process is the mixer. This device must be a nonlinear device that converts power from the original frequencies to the beat frequency, and is typically a diode because it

is a good approximation to a square-law device in terms of its current-voltage (I–V) behavior. If $I = V^2$, then the output current is proportional to the input power P , because V^2 is proportional to the electric field E^2 which is a measure of P . Within the mixer, the electric fields sum as vectors and the resultant power is the square of the amplitude of the electric fields. One of the best modern mixers is a junction made by separating two *superconductors* (not semiconductors) with a thin insulator; this is the SIS mixer.

The simplest radiometer (photometer) measures the average total power received over a well-defined radio frequency bandwidth $\Delta\nu$ and over a time interval τ . Just as for optical and infrared wavelengths, the weak astronomical source is measured against a background of many other radio signals such as the cosmic microwave background, the atmosphere, and the noise in the receiver itself. Power is usually expressed as a temperature, and the total system power is T_{sys} . The total noise in the measurement is given by the practical form of the radiometer equation:

$$\sigma_T = T_{\text{sys}} \left[(1/\Delta\nu_{\text{RF}}\tau) + (\Delta G/G)^2 \right]^{1/2} \quad (12.37)$$

where ΔG represents possible fluctuations in gain. If those fluctuations are negligible, then the equation simplifies to its ideal form ($T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$). In a manner similar to chopping in the infrared, one way to minimize fluctuations in receiver gain and atmospheric emission is to perform differential measurements by switching rapidly between two adjacent feeds as first suggested by Robert Dicke in the 1940s. The main drawback of Dicke switching is that the measured noise is doubled to $2 T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$ as a result of the difference measurement.

No quantum detector that uses the liberation of a (photo)electron by an incident photon can work at wavelengths longer than about 0.2 mm. Across the entire radio spectrum, the electromagnetic field, or the current which it induces in an antenna, is applied to a nonlinear element (diode) or mixer. The mixer either measures the total power or changes the signal frequency to one which is more easily measured. Frequently used devices are the Schottky diode and the superconducting junction (Zmuidzinas and Richards 2004).

When a metal and a semiconductor are brought into contact, the majority charge carriers of the semiconductor leave the contact zone until the Fermi levels of the metal and semiconductor are equalized. As a result, a “barrier” or “depletion region” empty of majority carriers appears in the semiconductor; typical barrier widths are $\sim 0.3 \mu\text{m}$. Even without any voltage across the junction, a current can flow through it. This is the Schottky diode. In practice, the semiconductor used is heavily doped gallium arsenide (GaAs) with a very thin lightly doped layer on the surface to reduce quantum mechanical tunneling and a contact layer of metal on top. Cooling the detector to a temperature of about 20 K yields a low-noise detector. Above 300 GHz (wavelengths shorter than 1 mm), the capacitance of the diode creates an RC filter which reduces its response.

When a thin insulating barrier is created between a normal metal and a superconducting metal (*not* a semiconductor) or between two superconductors, the structures are called SIN and SIS junctions. Nonlinear current can flow by quantum mechanical tunneling, and hence the device can be used as a mixer. When the voltage V is large enough that occupied states on one side are opposite vacant states on the other side, then a tunneling current can flow. Conversely, no current flows if $eV < 2\Delta$. Absorption of a photon can excite a charge carrier to the energy where the tunnel effect occurs. The high-frequency limit is about the same as the Schottky diode.

The output of the mixer is a signal with a frequency of $\nu_{\text{IF}} = \pm(\nu - \nu_{\text{LO}})$, where ν_{LO} is the frequency of the local oscillator, which is passed to the intermediate frequency (IF) amplifier and

then to a rectifying and smoothing circuit called a “detector.” This terminology seems strange compared to our previous usage where the detector is the device that receives the photons and converts them to an electrical signal. The detector part of the radio receiver is usually called the “backend” and can be quite complicated with a number of options, even for the same “front end” system. Rather than a single detector, a “backend” spectrometer often receives the output from the IF amplifier. The “spectrometer” usually consists of a number of electrical filters tuned to different frequencies with detectors on their outputs, a digital correlation computer.

4 Cryogenics and Vacuum Systems

Many detectors require cooling for optimum performance. There are several categories of cooling systems that might be required in astronomical instruments:

1. *Thermo-electric coolers and liquid circulation coolers.* These systems normally operate over the range -20°C to -50°C and are suitable for photomultiplier tubes, certain CCDs which have low dark currents, and high-speed applications such as telescope-guiding cameras.
2. *Liquid and solid cryogenics.* Dry ice (solid CO_2) is cheap and readily available. Coming in the form of a “snow,” it is most often used as the coolant for GaAs PMTs. Temperatures around -76°C (or 197 K) can be achieved with dry ice. Liquid nitrogen (LN_2) is also relatively cheap and can cool detectors (and other components) to -196°C (77 K), which is the normal boiling point of liquid nitrogen. Liquid helium (LHe) is more expensive, but it is needed to cool the low band gap semiconductor detectors and bolometers used in infrared instruments to -269°C (4 K). For liquid cryogenics, the cooling ability is expressed in terms of the product of the density (ρ) and the latent heat of vaporization (L_V):

$$\begin{aligned}(\rho L_V)_{\text{LHe}} &= 0.74 \text{ W hr l}^{-1} \\(\rho L_V)_{\text{LN}_2} &= 44.7 \text{ W hr l}^{-1}\end{aligned}$$

For example, a 10 W heat load boils away 1 liter (l) of LHe in 0.07 h, whereas 1 l of LN_2 lasts 4.5 h. By attaching a vacuum pump to the LN_2 vent and reducing the pressure above the liquid, it is possible to solidify the nitrogen and achieve ~ 65 K.

3. *Electrical heat engines or closed-cycle refrigerators.* Because LHe is fairly expensive compared to liquid nitrogen and requires special care in handling, many infrared instruments and submillimeter/radio receivers employ multistage closed-cycle refrigerators (CCRs). Typical two-stage Gifford-McMahon systems using 99.999% pure helium gas at about 300 psi as the working fluid, such as the 350 model from CTI Inc. (now part of Brooks Automation Inc., USA), can provide two cold levels, usually 65 and 10 K, and extract heat at a rate of about 20 and 7 W from each stage. Both single-stage and triple-stage versions are available, and the larger units provide about 100 W of cooling power. If the mass to be cooled is very small, for example, a single CCD, then much simpler systems such as small Stirling cycle coolers can be used. Vibration damping is needed, especially for CCR-cooled instruments in sensitive AO systems, and counterbalance weights may also be required.
4. *^3He systems.* Even lower temperatures are obtained using ^3He systems in submillimeter and far-infrared bolometers. The basic principle of a ^3He cryostat is to condense helium-3 gas by bringing it in contact with a pumped helium-4 reservoir (yielding ~ 1.2 K). Low temperatures below 300 mK are then obtained (for small samples) by reducing the vapor pressure on top of the liquid helium-3 using an internal sorption pump (charcoal).

The heat H (J) removed from a mass m (kg) which is cooled from a temperature T_h to T_c is given by

$$H = mC(T_h - T_c) \quad (12.38)$$

where C is the specific heat of the material in joule/kg per K ($\text{J kg}^{-1}\text{K}^{-1}$). The specific heat of a substance usually changes with temperature, and it depends on the conditions under which the heat is applied (i.e., constant volume C_v or constant pressure C_p), but for solids the difference is generally small. For aluminum, C is about 900 J/kg-K, copper is 385 J/kg-K, steel is about 450 J/kg-K, and water is 4,190 J/kg-K. Specific heat is also tabulated in cal/g-K; using the conversion 4.19 joules per calorie gives 1 cal/g-K for water. As an example, to cool a mass of 1 kg (2.2 lb) of copper from 290 K to 80 K requires the removal of $1 \times 385 \times 210 = 80,850$ J and 2.3 times that amount for aluminum. If heat removal is to be accomplished in $t = 1$ h (3,600 s), then the average power is $H/t = 22.5$ W.

The rate of transfer of heat Q_H (in watts) by conduction along a rod of uniform cross-sectional area (A) and temperature gradient dT/dx is given by

$$Q_H = -kA \frac{dT}{dx} \quad (12.39)$$

where k is called the thermal conductivity ($\text{W m}^{-1}\text{K}^{-1}$) and is about 240 for Al, 400 for Cu, and only 0.9 for glass. In steady state conditions, we can write dT/dx as $\Delta T/L$, where L is the length of the conductor. If we think of heat flow as “current” and temperature difference as “voltage,” then by analogy with Ohm’s law ($V = IR$) for electrical circuits, we can define a “thermal resistance” such that

$$\Delta T = Q_H R, R = \frac{1}{k} \frac{L}{A} \quad (12.40)$$

Since k is a function of temperature, it is often more convenient to integrate over the required temperature range and give Q_H in the form

$$Q_H = \frac{A}{L} [I_{T_h} - I_{T_c}] \quad (12.41)$$

where L is the total length of the conductor and I is a tabulated property for many materials called the thermal conductivity integral which accounts for the variation of thermal conductivity (k) with temperature. In this expression, T_h and T_c represent the hot and cold temperatures between which the heat is flowing. If A and L are measured in cm^2 and cm, respectively, then I is in W/cm. Typical values are shown in [Table 12-1](#).

In large cryogenic infrared instruments, heat transfer by radiation is the dominant mechanism. The power radiated from a body of area A at an absolute temperature T is given by

$$Q_H = \varepsilon \sigma A T^4 \quad (12.42)$$

where $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ is the Stefan-Boltzmann constant and ε is the emissivity of the surface; $\varepsilon = 1$ for a perfectly black surface and is less than 0.1 for a shiny metallic surface. Polished aluminum foil yields an emissivity of about 0.05 and goldized Kapton gives ~ 0.04 ; anodizing increases the emissivity by a factor of 10 or more. The net rate of heat transfer by radiation from a body at temperature T_h onto a body at temperature T_c is given by

$$Q_H = \sigma A_c F_{hc} [T_h^4 - T_c^4] \quad (12.43)$$

Here, F_{hc} is an “effective” emissivity which also depends on the geometry of the cryostat (or dewar) configuration, such as concentric cylinders or plane-parallel plates, both of which are

■ Table 12-1

Values of the thermal conductivity integrals in W/cm for four materials

Temperature (K)	OFHC copper (W/cm)	6061 aluminum (W/cm)	Stainless steel (W/cm)	G-10 fiberglass (W/cm)
300	1,520	613	30.6	1.00
250	1,320	513	23.4	0.78
200	1,120	413	16.6	0.55
150	915	313	10.6	0.37
100	700	211	6.3	0.19
77	586	158	3.2	0.11
50	426	89.5	1.4	0.07
10	25	3.6	0.03	0.005

quite realistic. In both cases, when the emissivities of the two surfaces are small (<5%) and equal, then $F_{hc} \sim \varepsilon/2$.

For a given temperature differential, radiation load is minimized by reducing the surface area and achieving the lowest emissivity (shiniest) surfaces. It is also possible to add n “floating” shields which reduce the radiated heat load on the innermost body by a factor of $(n + 1)$, but careful application is critical. Various forms of floating shields are available. One form is called Multiple Layer Insulation (MLI). Typical emissivities range from 0.03 for polished aluminum and gold foil to 0.32 for polished anodized aluminum to 0.95 for a matt black surface. A useful rule of thumb is that if $\varepsilon = 0.05$ and the hot and cold temperatures are 290 and 80 K, respectively, then the radiation load is $\sim 10\text{--}11 \text{ W/m}^2$. Note that radiation load is very sensitive to T_h .

Air at 20°C and 1 atmosphere of pressure (1 atmosphere = 760 torr and 1 torr = 132 Pa) contains about 2.7×10^{19} molecules per cubic centimeter, and the average distance traveled between collisions, called the mean free path, is about 7×10^{-6} cm. In general,

$$\lambda_{mfp} = \frac{1}{\sqrt{2}n\pi d^2} \quad (12.44)$$

where n is the number density of molecules and d is the diameter of the molecule. A “rough” vacuum is about 10^{-3} torr (mean free path = 5 cm) and a “high” vacuum would be 10^{-6} torr (mean free path 5×10^3 cm). The capacity of a vacuum pump is specified in terms of the pumping speed at the inlet, which is just the volume rate of flow $S = dV/dt$ l/s. The throughput of the flow is $Q_p = PS$ (torr-l/s) and the throughput of the pumping line is given by $Q_p = C\Delta P$, where ΔP is the pressure gradient and C is called the conductance which depends on gas pressure and viscosity. Since this equation is also analogous to Ohm’s law ($V = IR$) for electrical circuits, the net pumping speed of a pump and a system of pumping lines is given by

$$\frac{1}{S} = \frac{1}{S_{\text{pump}}} + \frac{1}{C_{\text{lines}}} \quad (12.45)$$

where the net conductance is found by adding the individual conductances like their electrical counterparts. Two equations are given for C (in l/s). The first corresponds to viscous flow when the mean free path is small and the other to molecular flow when the mean free path is large compared to tube dimensions and C is independent of pressure. Both apply to air at 20°C:

$$C = 180 \frac{D^4}{L} P_{av} \text{ or } C = 12 \frac{D^3}{L} \quad (12.46)$$

It is assumed that the tube is circular with diameter D and length L in cm and the pressure is in torr. Finally, the pump down time (in seconds) of a system with volume V from pressure P_0 to P assuming a constant net pumping speed S and no outgassing is

$$t = 2.3 \frac{V}{S} \ln \left(\frac{P_0}{P} \right) \quad (12.47)$$

Typically, the chamber is rough pumped to about 5×10^{-2} torr with a mechanical pump and then pumped to a lower pressure with a diffusion pump or turbomolecular pump. Typical pump speeds are 100 l/s at the inlet.

Developing modern astronomical instrumentation requires a broad knowledge of basic physics, engineering practice, and software. This chapter has touched on some of the background knowledge needed to appreciate what goes into the designing and building of such instruments. The design specification flows down from the scientific requirements. Initial optical, mechanical, and thermal calculations need to be followed by detailed ray tracing and finite element analysis. Many instruments require vacuum cryogenic systems. Analog and digital electronics are needed to operate the instrument and detector, and software to control the entire system must be robust and easy to use. Many other factors come into play when building astronomical instruments. See McLean (2008) for a more complete review of astronomical instruments and detectors.

References

- Allington-Smith, J., & Content, R. 1998, *PASP*, 110, 1216
- Arribas, S., et al. 1998, *Proc. SPIE*, 3355, 821
- Bacon, R., et al. 1995, *A&AS*, 113, 347
- Baldry, I. K., Bland-Hawthorn, J., & Robertson, J. G. 2004, Volume phase holographic gratings: polarization properties and diffraction efficiency. *Pub. Astron. Soc. Pac.*, 116, 403–414
- Barden, S. C., Arns, J. A., Colburn, W. S., & Williams, J. B. 2000, Volume-phase holographic gratings and the efficiency of three simple volume-phase holographic gratings. *Pub. Astron. Soc. Pac.*, 112, 809–820
- Bland-Hawthorn, J., & Kedziora-Chudczer, L. 2003, Taurus tunable filter: seven years of observing. *Pub. Astron. Soc. Australia*, 20, 242–251
- Clarke, D., & Grainger, J. 1971, *Polarized Light and Optical Measurement* (Oxford/New York: Pergamon Press)
- Courtes, G. 1982, *ASSL*, 92, 123
- Garcia, A. A., Rasilla, J. L., Arribas, S., & Mediavilla, E. 1994, *Proc. SPIE*, 2198, 75
- Garmire, G. P., Bautz, M. W., Ford, P. G., Nousek, J. A., & Ricker, G. R., Jr. 2003, *Proc. SPIE*, 4851, 28. X-ray and gamma-ray telescopes and instruments for astronomy, ed. J. E. Trümper, & H. D. Tananbaum
- Gehrels, N., Chincarini, G., Giommi, P., Mason, K. O., Nousek, J. A., Wells, A. A., White, N. E., Barthelmy, S. D., Burrows, D. N., Cominsky, L. R., et al. 2004, The *Swift* gamma-ray burst mission. *ApJ*, 611, 1005–1020
- Guyon, O. 2006, Theoretical limits on extrasolar planet detection with coronagraphs. *Astrophys. J. Suppl. Ser.*, 167, 81–99
- Hinkley, S., Oppenheimer, B. R., Brenner, D., Parry, I. R., Sivaramakrishnan, A., Soummer, R., & King, D. 2008, *Proc. SPIE*, 7015, 32
- Janesick, J. R. 2001, *Scientific Charge-Coupled Devices* (Bellingham: SPIE)
- Kasdin, N. J., et al. 2003, Extrasolar planet finding via optimal apodized-pupil and shaped-pupil coronagraphs. *Astrophys. J.*, 582, 1147–1161
- Krabbe, A., Thatte, N. A., Kroker, H., Tacconi-Garman, L. E., & Tecza, M. 1997, *Proc. SPIE*, 2871, 1179
- Larkin, J., et al. 2006, *Proc. SPIE*, 6269, 42
- Lyot, B. 1939, A study of the solar corona and prominences without eclipses. *Mon. Not. R. Astron. Soc.*, 99, 580–594
- McLean, I. S. 2008, *Electronic Imaging in Astronomy: Detectors and Instrumentation* (2nd ed.; Heidelberg: Springer)
- Morrissey, P., Schiminovich, D., Barlow, T. A., Martin, D. C., Blakkolb, B., Conrow, T.,

- Cooke, B., Erickson, K., Fanson, J., Friedman, P. G., et al. 2005, The on-orbit performance of the *Galaxy Evolution Explorer*. *ApJ*, 619, L7–L10
- Murphy, T. W., Matthews, K., & Soifer, B. T. 1999, *PASP*, 111, 1176
- Palacios, D. M. 2005, An optical vortex coronagraph. *Proc. SPIE*, 5905, 196
- Rieke, G.H. 2003, *The Measurement of Light from the UV to the Sub-millimeter* (Cambridge: Cambridge University Press)
- Rouan, D., et al. 2000, The four quadrant phase mask coronagraph. I. Principle. *PASP*, 112, 1479–1486
- Soummer, R., Aime, C., & Falloon, P. E. 2003, Stellar coronagraphy with prolate apodized circular apertures. *A&A*, 397, 1161–1172
- Tecza, M., Thatte, N. A., Krabbe, A., & Tacconi-Garman, L. E. 1998, *Proc. SPIE*, 3354, 394
- Tinbergen, J. 1996, *Astronomical Polarimetry* (Cambridge, UK: Cambridge University Press)
- Siegmund, O., Vallerger, J., Tremsin, A., and McPhate, J. 2007, Microchannel plates: recent advances in performance, *Proc. SPIE*, 6686, 66860W
- Wizinowich, P., Acton, D. S., Shelton, C., Stomski, P., Gathright, J., Ho, K., Lupton, W., Tsubota, K., Lai, O., Max, C., Brase, J., An, J., Avicola, K., Olivier, S., Gavel, D., Macintosh, B., Ghez, A., & Larkin, J., 2000, First Light Adaptive Optics Images from the Keck II Telescope: A New Era of High Angular Resolution Imagery, *PASP*, 112, 315
- Zmuidzinas, J., & Richards, P. L. 2004, Superconducting detectors and mixers for millimeter and submillimeter astrophysics. *Proc. IEEE*, 92, 1597–1616

13 Silicon-Based Image Sensors

Paul R. Jorden

e2v technologies plc, Chelmsford, Essex, UK

1	<i>Introduction</i>	543
1.1	Overview	543
1.2	Use of CCD and Silicon Sensors	544
2	<i>The Charge-Coupled Device Image Sensor</i>	545
2.1	Introduction	545
2.2	What Is a CCD?	545
2.3	Spectral Response	545
2.4	Readout Noise	549
2.5	The Electron-Multiplying (EM) CCD	551
2.6	Readout Frequency and Read Times	551
2.7	Dark Current and Operating Temperature	552
2.8	Charge Transfer Efficiency	553
2.9	Package Types	554
2.10	Single Devices and Mosaic Arrays	555
3	<i>CMOS/APS Imagers</i>	555
3.1	Introduction to APS Imagers	555
3.2	Why Use an APS Sensor?	558
3.3	Performance Features of APS Devices	558
4	<i>Sensor Types and Application Examples</i>	558
4.1	CCD Formats	558
4.2	Custom Spectral Response	559
4.2.1	Sensor Optimization for Different Bands	559
4.2.2	Graded AR Coating	559
4.2.3	Multilayer AR Coatings	559
4.2.4	X-ray and Short Wavelengths	560
4.3	Wavefront Sensors	560
5	<i>Summary and Future Trends</i>	560
5.1	Device Size	560
5.2	Pixel Size	562
5.3	Readout Noise	562

5.4	Spectral Response	562
5.5	Associated Electronics	563
	<i>References</i>	563

Abstract: Silicon image sensors are the dominant detector type used at visible wavelengths for astronomy. The primary charge-coupled-device (CCD) sensor is described with architecture, performance levels, and applications illustrated. CMOS silicon sensors are also discussed, with examples given of application areas. Most astronomical sensors require high quantum efficiency and low readout noise, and techniques to achieve these are presented. Silicon sensors are discussed for applications covering the x-ray through visible to near-infrared wavelengths. The special requirements of packaging for cryogenic use, close butting, and focal plane mosaic assemblies are also illustrated.

List of Abbreviations: *APS*, Active-pixel sensor; *AR*, Anti-reflection (coating); *CCD*, Charge-coupled device; *CMOS*, Complementary metal-oxide-semiconductor; *CTE*, Charge transfer efficiency; *ELT*, Extremely large telescope; *EMCCD*, Electron-multiplying CCD; *FPA* Focal plane assembly; *FPS*, Frames per second; *IMO*, Inverted mode operation; *MPP*, Multiphase pinned; *PGA*, Pin grid array; *Pixel*, Picture element; *QE*, Quantum efficiency; *WFS*, Wavefront sensor

1 Introduction

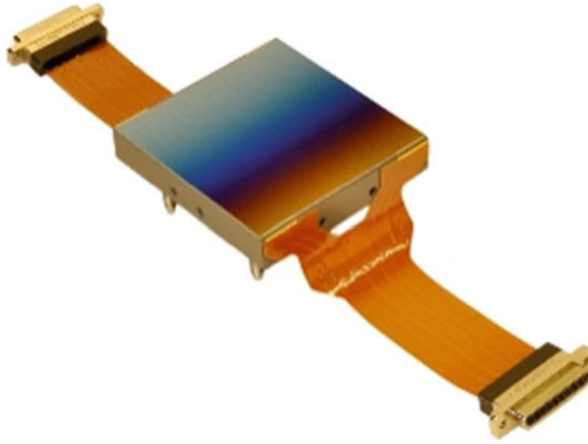
1.1 Overview

Astronomy research advances through a synergy of theoretical understanding and observational measurements. The main requirement for astronomical measurements is the achievement of high sensitivity for faint signals. Particularly in the last three decades, silicon sensors have been the dominant optical sensor used on telescopes. Extremely high sensitivity is available together with large focal plane collecting areas which together deliver highly efficient optical light detection. Historical information, technical details, and applications to instrumentation can be found in Janesick (2001) and McLean (2008).

In this chapter, the charge-coupled-device (CCD) characteristics are discussed. This device is a two-dimensional image array capable of providing a high-precision digital image. Optical sensitivity, wavelength range, readout noise, formats, and sizes are all important parameters that define their use in optical instruments. Many other features such as readout time and even device package design have an influence in astronomical instrument design and are also discussed here. Quantum efficiency (spectral sensitivity) and readout noise (minimum detectable signal) are probably the two most important operational parameters.

► *Figure 13-1* illustrates a typical modern charge-coupled-device sensor. It has a high optical sensitivity and very low noise, precision package for cryogenic focal plane mounting, available in multiple performance variants, and is the heart of many astronomical instruments; also see e2v (www.e2v.com/imaging) for examples of sensors.

A discussion of optical (silicon) sensors would not be complete without including the CMOS or APS sensor (see abbreviation list). These devices are very familiar to cell-phone camera users but have only recently started to be used for astronomy. The specific advantages and disadvantages of this type of sensor are discussed, particularly in comparison with the established and widely used CCD.



■ Fig. 13-1
The e2v CCD231-84, 4,096 × 4,096 pixel astronomical sensor

1.2 Use of CCD and Silicon Sensors

The CCD is primarily an image sensor and is often used in large focal planes for this purpose, including many major survey telescope applications. A typical intermediate-sized sensor could have $2,048 \times 2,048$ pixels (or 30×30 mm) for use as a single square-format imager. Larger devices are also available although the maximum size is limited by silicon wafer diameter which is 150 mm for most specialist CCD manufacturers. For those telescopes that use a large focal plane area (e.g., 500 mm diameter), the usual solution is to build a two-dimensional mosaic of close-butted CCDs.

The CCD also has a significant use as a spectroscopic sensor, with the ability to provide a large image format and very low readout noise. Particularly for spectroscopy, with signal levels reduced by dispersion, the attainment of very low readout noise is important. Echelle and multi-object spectrographs often benefit from large-format sensors, including close-butted mosaics, where appropriate.

High quantum efficiency is usually required and can be achieved by silicon sensors as discussed in following sections. The ability to detect small signals requires low readout noise – and this is also achieved. In astronomical imaging and spectroscopy, exposure times are often in the range 10–3,000 s, and so rapid data readout is not demanded. This allows modest pixel readout rates and helps to achieve low read noise.

Other auxiliary function application areas include guide cameras and wavefront sensors, and these are discussed and demonstrate the versatility of this sensor. In these cases, high frame rate is often more important than large device size.

While CMOS (or APS) sensors do not have the heritage of CCDs for astronomy, they are finding specific uses where their features offer advantages, and these are discussed.

Many space instruments utilize CCDs in the traditional “ground” wavelength ranges and particularly in regions where atmospheric absorption favors space telescopes. CCD use in the x-ray, and short UV wavelengths are discussed together with visible wavelengths.

Finally, some indications of performance limitations and future directions are discussed.

2 The Charge-Coupled Device Image Sensor

2.1 Introduction

The charge-coupled device (CCD) is a silicon circuit designed for use as a high-performance image sensor. The basic device delivers a two-dimensional electronic image that would be stored and analyzed by a linked electronic and host computer system. The most direct application is to record images (normally of defined wavelength), although the same sensor is widely used for spectrographic recording as well as more specialized image sensor uses.

The silicon integrated circuit is familiar as a microcircuit component for laptops, mobile phones, and many other essential items of the modern world. Standard integrated circuits are traditionally built for high operating frequency, low power, and low cost – which usually means that they are designed with the smallest internal dimensions possible. In contrast, the CCD image sensor for astronomical applications is designed for the highest light detection efficiency, wide wavelength range, and large spatial format. The CCD considerably exceeds the sensitivity of the photographic plate and has the huge benefit of direct electrical signal output and well-calibrated response. Here, we discuss the design and performance features of the CCD that is designed for use as an astronomical light sensor.

2.2 What Is a CCD?

The standard design of a CCD consists of a two-dimensional matrix of light-sensing elements (pixels) that are coupled to a readout circuit. Incident photons strike the silicon and are converted into electrical charge. The specific architecture of devices can include variations of this structure, depending on the required application. The figure below illustrates a typical CCD architecture (▶ [Fig. 13-2](#)).

A CCD is fabricated using multiple layers of polysilicon gates which create the pixel structures. The device, in its simplest form, is operated as a frontside-illuminated device – whereby incident light passes through the semitransparent polysilicon layers.

However, the polysilicon absorbs some of the light, and this limits the total efficiency as well as cutting off the short wavelength sensitivity. For demanding scientific applications like astronomy, so-called “backside-illuminated” CCDs are almost universally used. In this case, the underlying silicon substrate of the device is removed, and the device is inverted so that incident light strikes the back surface. Charge is now collected very efficiently. ▶ [Figure 13-3](#) illustrates this mode of use.

Incident photons are absorbed and generate charge which needs to be collected and then reach the buried channel region – where it can be transferred to the outputs and measured.

2.3 Spectral Response

The most important reason for the dominant use of CCDs in astronomy is their very high quantum efficiency. Silicon can absorb photons very efficiently and achieve almost 100% quantum efficiency at some wavelengths. Quantum efficiency is the efficiency with which incident

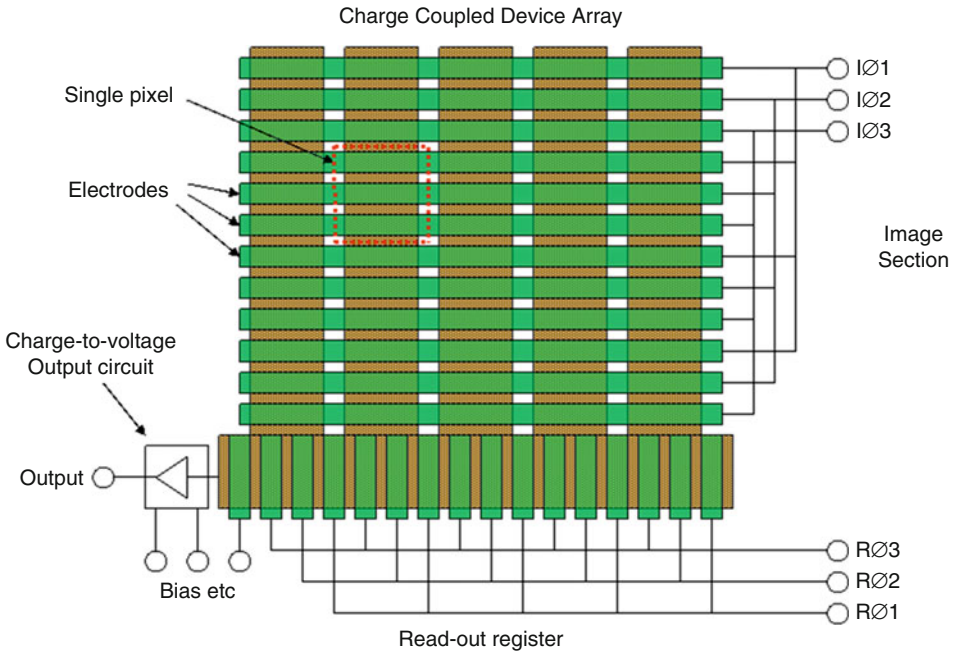


Fig. 13-2 Architecture of a typical scientific CCD

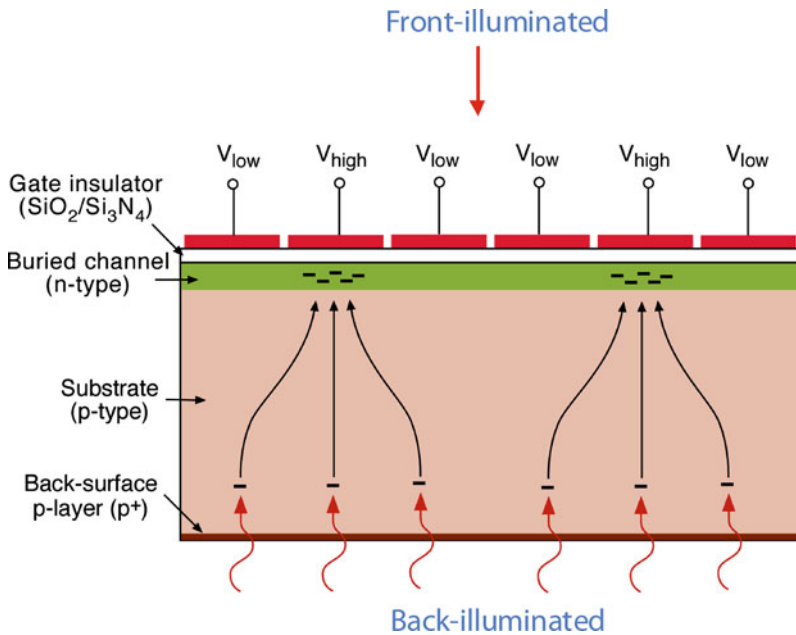


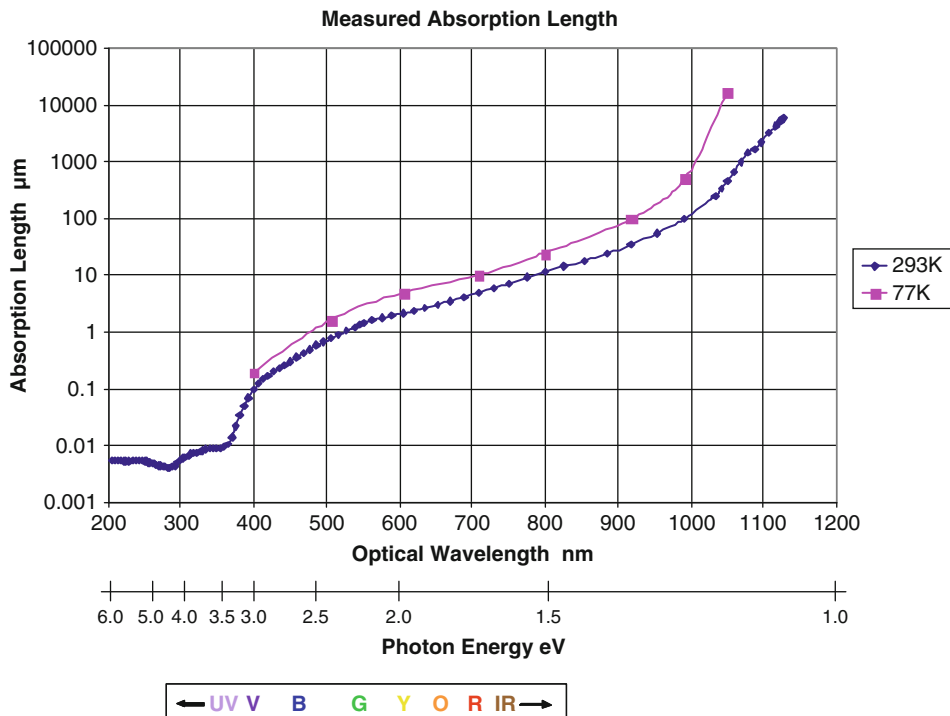
Fig. 13-3 Light detection by frontside- and backside-illuminated CCD

photons are converted to charge; in an ideal case, each incident photon generates one electron that would be collected. The backside-illuminated device (as discussed above) allows high-efficiency sensors.

Silicon has a variable absorption coefficient which determines the efficiency with which differing wavelengths can be absorbed. The figure below shows how absorption length varies with wavelength. There is also some temperature dependence, which can be relevant since many astronomical detectors are operated at cryogenic temperatures (e.g., -100°C) (► Fig. 13-4).

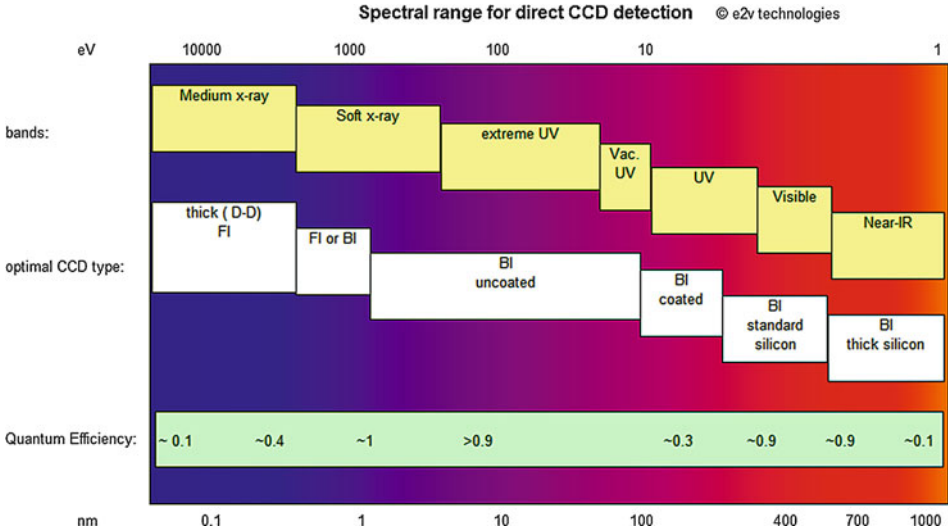
The varying absorption depth has several consequences for the detection of light. Short wavelengths (UV) are absorbed in very shallow depths, and such signal can be lost before it reaches the active (charge collection) region of a sensor; this means that back-illuminated sensors with minimal “dead” layers near the surface are required. Midrange wavelengths (visible) penetrate and are collected with a depth of about $1\ \mu\text{m}$ and give high QE. Long wavelengths ($>800\ \text{nm}$) require 10 s of microns of silicon for high absorption, and this requires particularly thick silicon sensors. Silicon has a band gap of about $1.1\ \text{eV}$ which prevents detection of photons beyond about $1,100\ \text{nm}$.

In fact, silicon sensors can be used to directly detect photons over a wide range of wavelengths/energies, covering the x-ray to near-infrared range of the electromagnetic spectrum, as shown in ► Fig. 13-5. The quantum efficiency varies as the absorption depth varies at differing photon energies.



► Fig. 13-4

Silicon absorption length



■ Fig. 13-5

Spectral range of silicon photon absorption

Silicon has a refractive index of about 4 at visible wavelengths and therefore can reflect approximately 35% of the incident light. It is a standard practice to apply an antireflection (AR) coating to the CCD to minimize this loss. The result is that over 90% efficiency can be obtained. The AR coating can be varied to give different peak QE values at chosen wavelengths.

As discussed above, the silicon thickness has a strong influence on detection efficiency at longer wavelengths. Standard silicon devices are made of low-resistivity silicon which has a depletion depth of order 10 μm ; this means that devices need to be manufactured with this thickness for effective charge collection, resulting in a modest or poor QE at long wavelengths. Increasingly higher resistivity silicon is now used to allow thicker devices with increased red wavelength QE.

The figures below show examples of spectral response for different CCD types.

► *Figure 13-6* shows a typical spectral response of a “deep depletion” device with modern multilayer AR coating, optimized for the 400–900 nm range. The variation of red sensitivity with operating temperature is indicated; in most cases, astronomical sensors require cryogenic cooling (for low dark signal), and this has the effect of reducing the red wavelength sensitivity.

► *Figure 13-7* shows devices of three different silicon thicknesses, all with an AR coating optimized for the 500–1,000 nm range. This illustrates the benefit of increasing device thickness at red wavelengths; “deep depletion,” “bulk,” and “high-rho” device types correspond to increasing silicon depth with nominal thickness of 40, 100, or 200 μm respectively.

CCD QE data for nonvisible wavelengths has been presented by Stern (2004) and Turner et al. (2001). Development of specialized thick silicon high-rho devices has been described by Holland, et al. (2007) and Jorden et al. (2010).

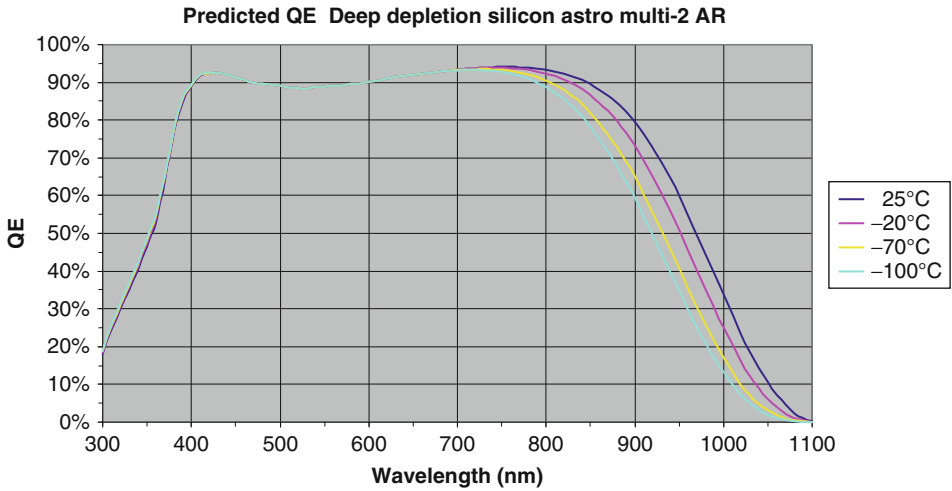


Fig. 13-6
"Visible" QE (400–900 nm)

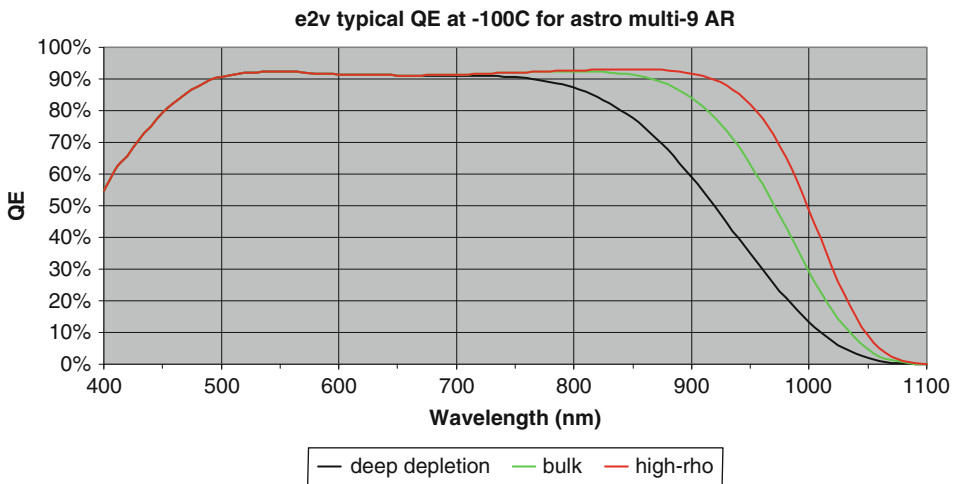
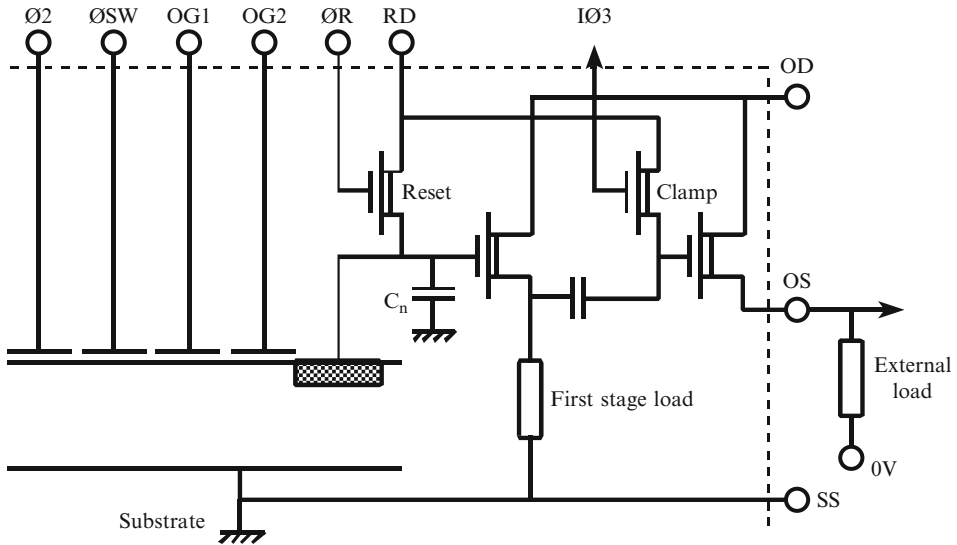


Fig. 13-7
IR QE for different device types

2.4 Readout Noise

In addition to extremely high quantum efficiency, CCDs are capable of measuring very small electrical signals. This combination of high QE and low noise allows high signal/noise measurements to be recorded. Incident photons are converted to electrical charge when detected and then converted to a voltage at an output node of the device, represented by the node capacitance C_n in the [Fig. 13-8](#).



■ Fig. 13-8
Example of low-noise CCD output circuit

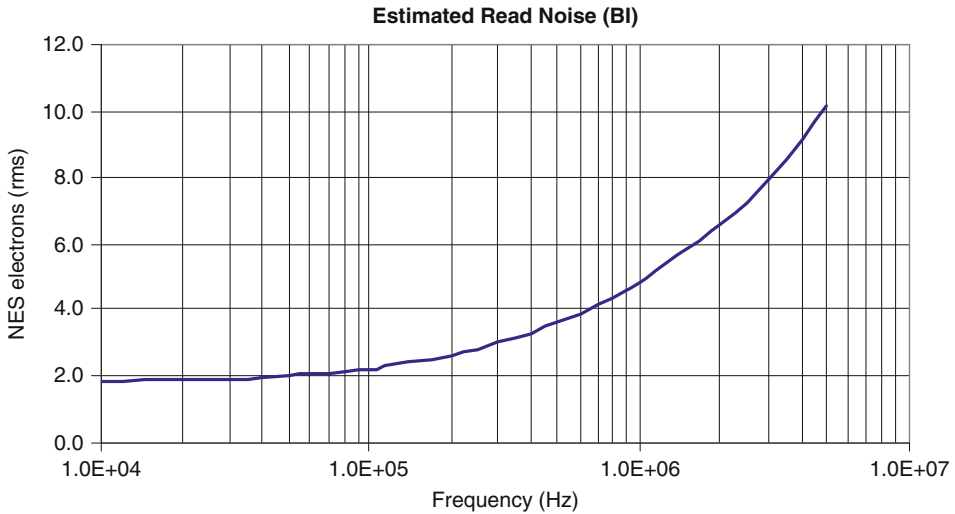
The signal is buffered by an output transistor prior to external measurement of the signal level. Other components of the circuit shown include a second-stage transistor buffer for improved external drive capability, and reset transistors used to establish a DC signal level as part of the signal sampling chain.

Output amplifiers are key components of the signal measurement process and may be designed for various optimizations. For astronomical applications, it is usual to utilize a very low-noise output amplifier. The read noise is defined as the standard deviation (1 sigma) of the output signal when no signal charge is present. The noise is a function of readout frequency, and the figure below illustrates typical performance of a high-performance CCD, with a read-noise floor of approximately $2 e^-$ rms at low frequencies (► Fig. 13-9).

The lowest noise is obtained at low frequencies, and for very large arrays, this requires a fairly long read time for the whole array; e.g., 160 s for a $4 K \times 4 K$ sensor, with one output at 100 kHz pixel rate. In some applications, higher frequencies and multiple outputs are used in order to reduce total read time.

The above examples refer to the commonly used “buried channel” output transistor. This has the lowest noise and is normally used for astronomical imaging. It is possible to use “surface channel” transistors which can operate at higher frequencies (e.g., 40 MHz); however, these have a higher characteristic noise and are not normally used for low-level imaging applications.

Amplifiers that are designed for lowest noise usually require small transistors which have limited drive capability and hence poor ability to operate at higher frequencies. Designing for higher speed operation leads to devices with higher read noise; the electron-multiplying CCD (see next section) overcomes this problem by providing subelectron equivalent read noise even at higher frequencies.



■ Fig. 13-9
Typical read noise versus frequency

2.5 The Electron-Multiplying (EM) CCD

A simple CCD (as above) operates by directly transferring the photo-charge (electrons) to the output circuit where it is measured. The EMCCD includes an extra transfer register which introduces gain into the signal chain so that the electron signal is amplified before reaching the output. This means that the effect of amplifier noise is reduced inversely in proportion to the gain. Typical gain values are X1,000, which means that an amplifier noise of $100 e^-$ rms is equivalent to a photo-signal charge of $0.1 e^-$. With almost 100% quantum efficiency, this means that individual photons can be detected with this type of CCD. Furthermore, it allows operation at higher pixel frequencies since the level of amplifier noise is now less important.

➤ *Figure 13-10* illustrates the architecture of this device type.

The device operates through charge multiplication which adds stochastic noise to the signal. This means that at zero signal levels, the read noise can be sub-electron, but with finite signals the effective noise is normally $\sqrt{2} \times$ signal. The device requires a high-voltage clock (of order 50 V amplitude) which makes it somewhat more complex to operate than a simple CCD. See Mackay (2012) and Tulloch and Dhillon (2010) for use of such devices in astronomy.

2.6 Readout Frequency and Read Times

As shown above, CCDs read out their signal through a combination of parallel and serial charge movements. Signal measurement is made through output ports operating at pixel frequencies in the 10 kHz to 10 MHz frequency ranges for most applications. The total readout time is usually dominated by the time taken to sample all of the pixels; i.e., read time = total number of pixels ($X \times Y$) / (pixel frequency * number of outputs). As devices increase in size, it is increasingly valuable to use 4 or 16 outputs to achieve sufficiently low readout times. Furthermore, as

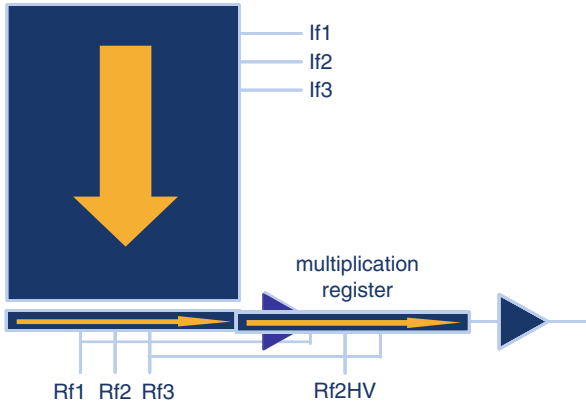


Fig. 13-10
Electron-multiplying CCD

shown in Sects. 2.4 and 2.5, modern devices offer very low readout noise even at pixel frequencies in the region of 1 MHz.

In some applications, such as guiders, devices are read out in windowed mode. In this case, the region of interest is small compared to the total frame size, and time to transfer charge from one row to the next has more impact on the total readout time.

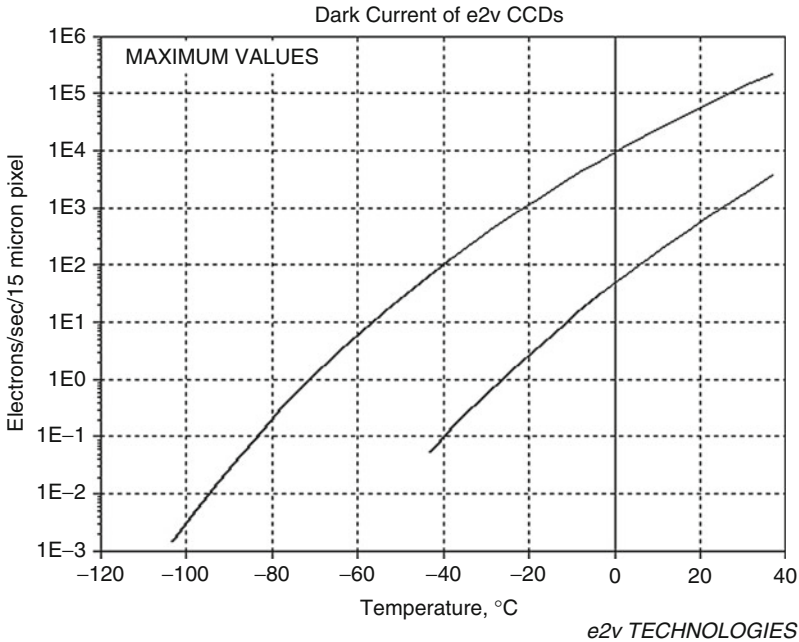
For other applications, such as large area wavefront sensing, it is necessary to read out a significant image size (e.g., $1,000 \times 1,000$ pixels) in times approaching 1 ms. This proves virtually impossible with a CCD since a high number of outputs operating at extreme pixel frequencies are required. In this case, the active-pixel sensor (or CMOS device) offers advantages, as discussed in Sect. 3 below.

2.7 Dark Current and Operating Temperature

As discussed above, a key feature of the CCD is its very low read noise. However, this benefit is only valuable if there are no additional sources of background signal. In particular, the silicon is capable of thermally generating charge known as dark current. The rate of charge generation is temperature dependent, and in order to allow long exposures with negligible dark current contribution, it is common to cryogenically cool astronomical CCDs to a temperature in the region of -100° .

In some cases, this degree of cooling is impractical, and it is possible to reduce the dark current by using so-called inverted mode operation (IMO). This is also known as “multiphase pinned” (MPP). Devices can be specifically made to allow this mode, with a modified clock waveform pattern, to reduce dark current by approximately a factor of 100. The disadvantage is smaller charge handling capacity and slower clocking speeds.

The figure below illustrates typical dark current for these two modes of use as a function of temperature (Fig. 13-11).



■ Fig. 13-11
CCD dark current versus temperature

2.8 Charge Transfer Efficiency

CCDs rely on transferring charge from one pixel to the next. This is done in parallel for the row transfer operations and in serial form for register readouts. It is usual to quote the charge transfer efficiency (CTE) of this operation in the form 99.999% or 0.99999%, normally quoted per pixel. It is essential that this efficiency is very close to unity, especially for the larger devices.

Fractional total charge after n transfers is $(S/S_0) = (C)^n$,

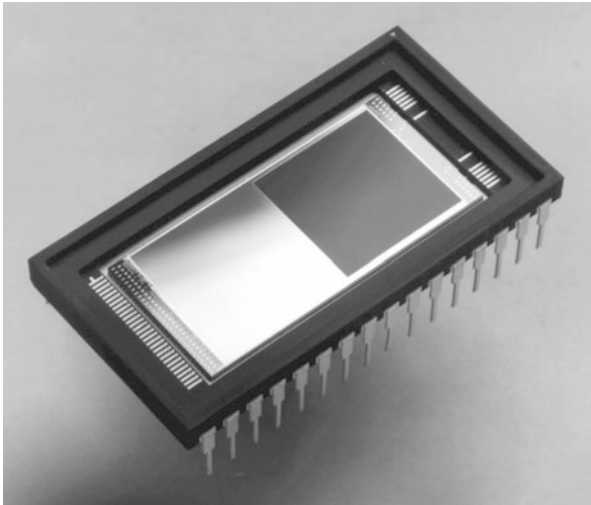
where S_0 = initial charge signal, S = final charge, C = CTE, and n = number of transfers (pixels).



For example, a device with 4,096 rows and $CTE = 0.99999$ will have 96% remaining charge after it is transferred from the top of the image to the readout register.

In reality, charge transfer loss is a stochastic process where charge is “lost” at multiple discrete locations as charge transfers through the device; the quoted CTE value is a mean value. In fact, the charge can be released again later with a characteristic time constant. Thus, CTE is a complex function of operating temperature, operating frequency, and signal level. A full description is beyond the scope of this discussion.

Astronomical CCDs are usually characterized by using a Fe^{55} x-ray source to generate a known signal level ($1,620 e^-$), with analysis software to determine lost charge. A typical CTE value of 99.995% can be expected. It is also possible to use a flat field image of known signal level; in this case, residual signal in overscanned pixels can indicate charge loss.

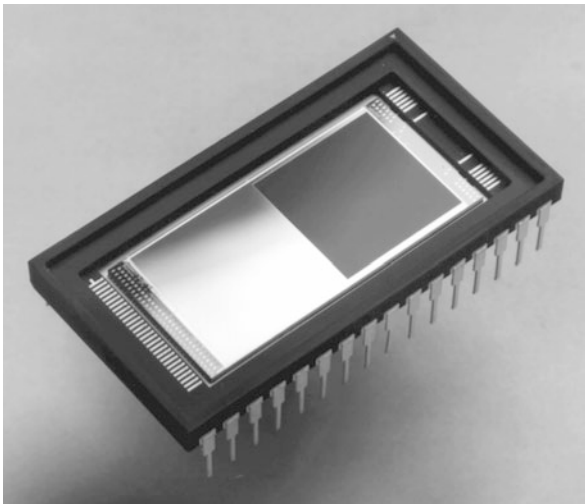
2.9 Package Types


Image sensors can be supplied in many different package types. A common style is the simple ceramic design which is widely used in the electronics industry. The ceramic base has internal circuit tracks which allow the silicon bond pads to be linked to pins; the ceramic package can be readily produced in quantity after initial design work and is therefore quite economic to manufacture. Pins can be arranged on two sides (dual-in-line) or underneath (pin grid array) formats.  *Figure 13-12* illustrates an example of the former type.

However, for mosaic focal planes (as in  *Sect. 2.10*), it is very desirable to have packages that have minimal edge gaps in order to achieve a high fill factor of the mosaic. In this case, two-, three-, or four-side buttable packages are required, as shown previously in  *Fig. 13-1*. In these cases, the device bond wires can be connected to flex cables with minimal lost spacing at the edge. Such devices are fragile to handle and require particular care when assembling close-packed mosaics.

Common types of package construction include:

Ceramic PGA or DIL	Simple package	Not suitable for close butting
Integrated Peltier	More specialized package	Requires sealed window to avoid condensation
Kovar/metal tub	Simple package	Not well suited for good flatness at cryo temperature
Metal package base	Can be customized and buttable	
Variant flexi packages	Flex cable facilitates assembly in mosaics	More complex/expensive



 **Fig. 13-12**

Example of ceramic package

Most silicon sensors are cooled below ambient in order to establish a low enough level of dark signal. This means that packages must be designed to allow for any expansion mismatch between the package material and silicon; stresses must be low enough to prevent damage and also to minimize distortion that can affect device focal surface flatness. Hybrid packages using metal and ceramic components together with flex cable connectors may be used for demanding applications.

Package materials include:

Material	Advantages	Disadvantages
Aluminium oxide (AlO)	Widely used, economical	Limited expansion match to silicon
Aluminium nitride (AlN)	Good expansion match to silicon	Slightly more expensive than AlO
Silicon carbide (SiC)	Stiff, low mass, good expansion match to silicon; high thermal conductivity	More expensive; specialized manufacture
Invar	Low expansion near ambient temperature, can be machined	Poor thermal conductivity; does not match silicon expansion at low temperatures
Kovar	Simple	Limited expansion match

2.10 Single Devices and Mosaic Arrays

For many applications, a single sensor is sufficient – especially since devices can now be manufactured with a 90×90 mm size and $10,000 \times 10,000$ pixels. The size of a single device is constrained by two main factors – silicon wafer size (which is normally 150 mm maximum diameter for specialized sensor manufacturers), and device manufacturing quality/yield (which is a strong function of device area). In practice this leads to devices of maximum size approximately 100 mm square or equivalent rectangular sizes. Examples of both are illustrated in [Fig. 13-13](#).

However, there are numerous telescopes and instruments that require a focal plane detection area of larger size, and these require multiple sensors usually constructed as a mosaic FPA (focal plane assembly). Impressive examples of large mosaics constructed for two space telescopes can be seen in the [Fig. 13-14](#); the *Kepler* and *GAIA* focal planes used custom sensors. For ground-based astronomy mosaics vary in size from pairs of sensors up to the much larger FPAs of Pan-STARRS and LSST. The reader is referred to the web pages for both projects. See Pan-starrs (<http://pan-starrs.ifa.hawaii.edu/public/design-features/cameras.html>) and LSST (<http://www.lsst.org/lsst/gallery/camera>) for further info.

Acknowledgements: *Kepler* (courtesy NASA & BATC) and *GAIA* (courtesy ESA & Astrium).

3 CMOS/APS Imagers

3.1 Introduction to APS Imagers

The CCD imager (as described above) has been widely used for astronomy. It has the advantage of a well-established heritage, very low readout noise, and is available in a variety of design types.

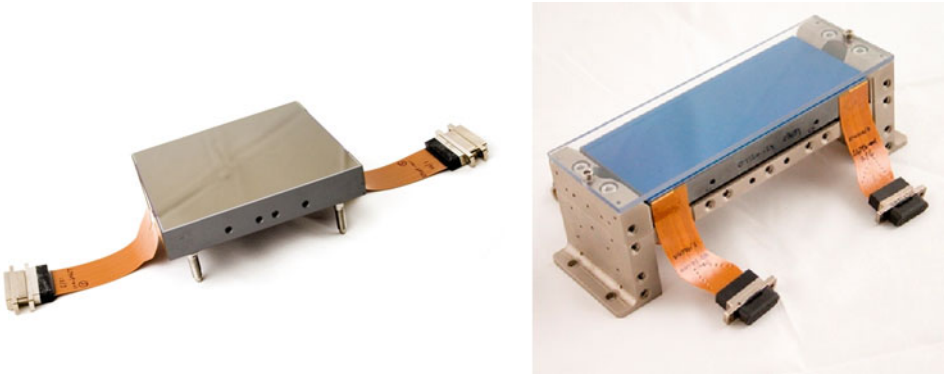


Fig. 13-13
92 × 92 mm CCD and 123 × 46 mm CCD



Fig. 13-14
(a) Kepler and (b) GAIA focal planes

The architecture relies on physically transferring charge internally until it reaches an output node where it is sensed. For many applications, the CCD remains the preferred sensor type.

However, the CMOS or active-pixel sensor (APS) does offer advantages for several applications. These devices are manufactured using CMOS (complementary metal-oxide-semiconductor) technology, which is extensively used for most integrated circuits such as

microcomputers, cell phones, etc. This technology allows a very high degree of integration with very small features and particularly allows high functionality within each pixel and within the device as a whole.

For scientific imaging applications, since high-performance image sensing is required, it is common to develop the devices specifically for this requirement and not necessarily integrate further functionality (such as digital image processing) into the device. Cell phones, on the other hand, are designed for maximum functional integration and minimal total size, and so their specifications differ from those of astronomical APS devices.

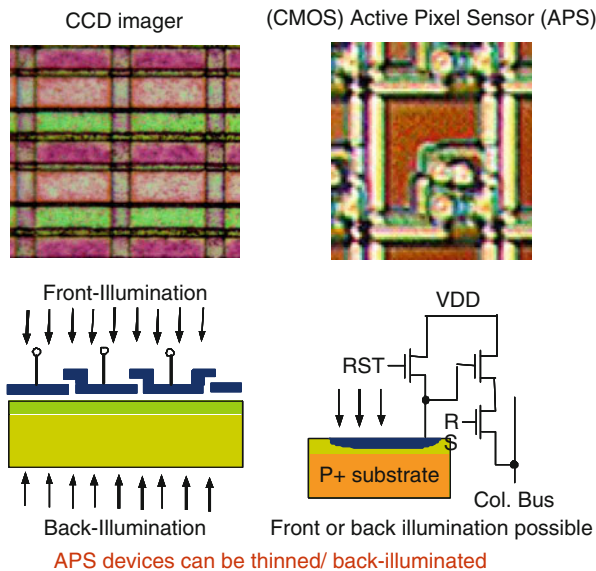
► *Figure 13-15* illustrates a CCD-imager pixel compared to an APS pixel.

Many of the performance features described in ► *Sect. 2* for CCDs apply also to APS sensors. The main differences relate to the architecture and readout method.

An APS sensor uses multiple transistors associated with each pixel to allow the signal to be measured and connected to circuitry at the edge of the device. Device designs can be characterized as 3T, 4T, 5T, etc., – referring to the number of transistors and the degree of internal functionality. Each pixel is accessed using a row/column address structure, resulting in multiple tracks across the surface of the imager. If the device is made in back-illuminated form (as for a CCD), then high quantum efficiency is achieved.

Because the pixels can be accessed in parallel, this allows faster readout of the signal than usually obtained for CCDs. Thus, the APS device is particularly suited for high frame rate applications such as high-speed photometry and wavefront sensing of large formats. Another significant advantage is very low power consumption; this has particular benefits for spaceborne applications, where power budget is often a constraint.

Since the signal from each pixel is independently sampled, this can give rise to larger pixel-pixel (sensitivity) variation than is usual for CCDs. However, careful device design and postcalibration generally prevent this from being a significant issue for astronomy.



■ Fig. 13-15

Comparison between CCD and APS sensors

3.2 Why Use an APS Sensor?

For many applications, such as large area “traditional” imaging, the CCD is adequate. In fact, the CCD has high performance and is often the best solution since it is well established, with well-understood performance and available in a variety of designs. The APS device has the advantage when higher speed readout and large area are required. The parallelism of the internal design allows high-speed readout to a degree not possible with CCDs.

This can be illustrated with the example of the wavefront sensor application for adaptive optics – which requires approximately 1,000 frames per second (fps) readout. A CCD with 256×256 pixels can achieve this rate with multiple outputs. However, when the frame size exceeds 512×512 pixels, an excessive number of outputs or excessive pixel frequency is needed. In this case, the APS can do the job, allowing 1,000 fps with parallel readout of many pixels using internal signal sampling and high bandwidth digital outputs.

3.3 Performance Features of APS Devices

APS devices can now approach the performance of CCDs, especially with regard to the key parameters of quantum efficiency and readout noise. The silicon can be back-thinned to give high quantum efficiency. However, the lower operating voltages and the necessity to use thinner silicon means that red wavelength response will be lower than for CCDs. APS devices can be made with very small sense node size and hence high responsivity (e^-/volt); this translates to achieving a low readout noise – values down to $1 e^-$ have been achieved. However, unlike CCDs, the performance of every pixel can differ, and so it is more important to consider the distribution of noise levels across all pixels to get a real estimate of the device performance. State-of-the-art devices can achieve noise levels in the $2\text{--}3 e^-$ rms noise level which is competitive with CCDs. There are other parameters that also need care to ensure good scientific performance, including linearity, signal lag, and pixel response uniformity.

4 Sensor Types and Application Examples

4.1 CCD Formats

There are several architectures that can be used for CCDs; these include, full frame, split full frame, frame transfer, split frame transfer and interline transfer. For astronomical imaging, the demand for high quantum efficiency normally precludes use of the interline transfer type since this has considerable lost detection area due to the interline transfer columns.

For direct science imaging (including spectroscopy), the full frame type is most commonly used. In this case, an external electromechanical shutter defines the exposure time, and read out of image signal is performed after the shutter has been closed. Exposure time is defined by the shutter, and readout time is defined by the pixel count, pixel frequency, and number of outputs (see [Sect. 2.6](#)).

A simple variant of the full frame array uses a midsection split so that the device can be read out through two registers; this allows a halving of the readout time at the expense of using double the number of output channels. This format halves the number of parallel transfers required which can minimize CTE losses, especially for large arrays.

For applications such as autoguiding or high-rate photometry, it is necessary to have a high continuous cadence of readouts. In this case, the frame-transfer architecture allows exposure to the “image” section, rapid transfer to the “store” section, and then repetition of this cycle. The image data can be read out from the store at the same time as the next image is exposed to the image section. No shutter needs to be used, and the store section has an opaque light shield. There is the potential for some image smear due to illumination falling on the image area while this is being transferred to the store; this is minimized if the ratio of FT shift time to image exposure time is minimized.

It is possible to increase readout speed further by using the split frame-transfer architecture. Here, the split format halves the readout time and also halves the frame transfer time, which has the benefit of reducing potential frame transfer image smear.


4.2 Custom Spectral Response

As described above, almost all astronomical sensors would be back-thinned and use an antireflection coating for maximum sensitivity. A single layer AR coat provides this to first order, but will not be perfect at all wavelengths. For applications requiring a wide spectral range on one sensor, then it is desirable to maximize the response; several examples are discussed.

4.2.1 Sensor Optimization for Different Bands

Some instruments divide the wavelength bands into different channels (e.g., multiband photometers or multiarmed spectrographs). In these cases, it is possible to optimize the sensitivity by implying a differently optimized sensor for each channel; usually each sensor is of the same type, but merely with different response characteristics. The simplest example would be to use a standard silicon CCD with blue/visible AR coating for a “blue” spectrograph arm, and a thicker deep depletion silicon CCD with red AR coating for a “red” arm.

4.2.2 Graded AR Coating

If a fixed format spectrum is projected onto the sensor, then it is possible to tune the AR coating in the spatial dimension to match the projected spectrum. This results in an optimized response at every wavelength. It is possible to design techniques to spatially vary the AR coating to match the spectrograph design, even with curved cross dispersed formats or nonlinear dispersion. The design is relatively insensitive to small degrees of mismatch and is generally a useful improvement over a simple coating.  *Figure 13-1*, as used in the introduction, also shows an example of a large format ($4\text{ K} \times 4\text{ K}$) sensor with such a coating.

4.2.3 Multilayer AR Coatings

In many cases, such as wide band imaging, the most useful optimization is to have the highest possible response across the whole sensor surface. In this case, multilayer coatings are valuable. Multilayer coatings are widely used in optical instruments on various glass (or similar)

components. Silicon sensors require specialized AR coatings for various reasons including high refractive index of silicon, safe deposition process that does not damage sensor electrical properties, reliable use in vacuum at cryogenic temperatures and materials that are appropriate for use with silicon sensor manufacturing processes.


Because CCDs in particular are well established as astronomical sensors, manufacturers are turning to second-order refinements such as these multilayer coatings. They can offer the advantages of wider spectral range and can particularly give improvements at the extremes of the spectral range (such as UV and near-IR) where signal/noise is often poorest.

An efficient AR coating, by definition, gives minimum reflectivity. This enhances spectral response but has the secondary benefit of reducing ghost image or other instrument reflections. Furthermore, at red wavelengths, a low reflectivity ensures that fringe magnitudes are minimized.

4.2.4 X-ray and Short Wavelengths

Silicon sensors have good response to UV wavelengths and x-rays. Good UV response requires high-performance back-thinning, with minimal dead layer thickness. This also applies to low-energy x-rays. Medium-energy x-rays are detected well with normal sensor thickness in the 10–40 μm range. At high energies, x-ray absorption efficiency falls and requires thick silicon for useful response.

4.3 Wavefront Sensors

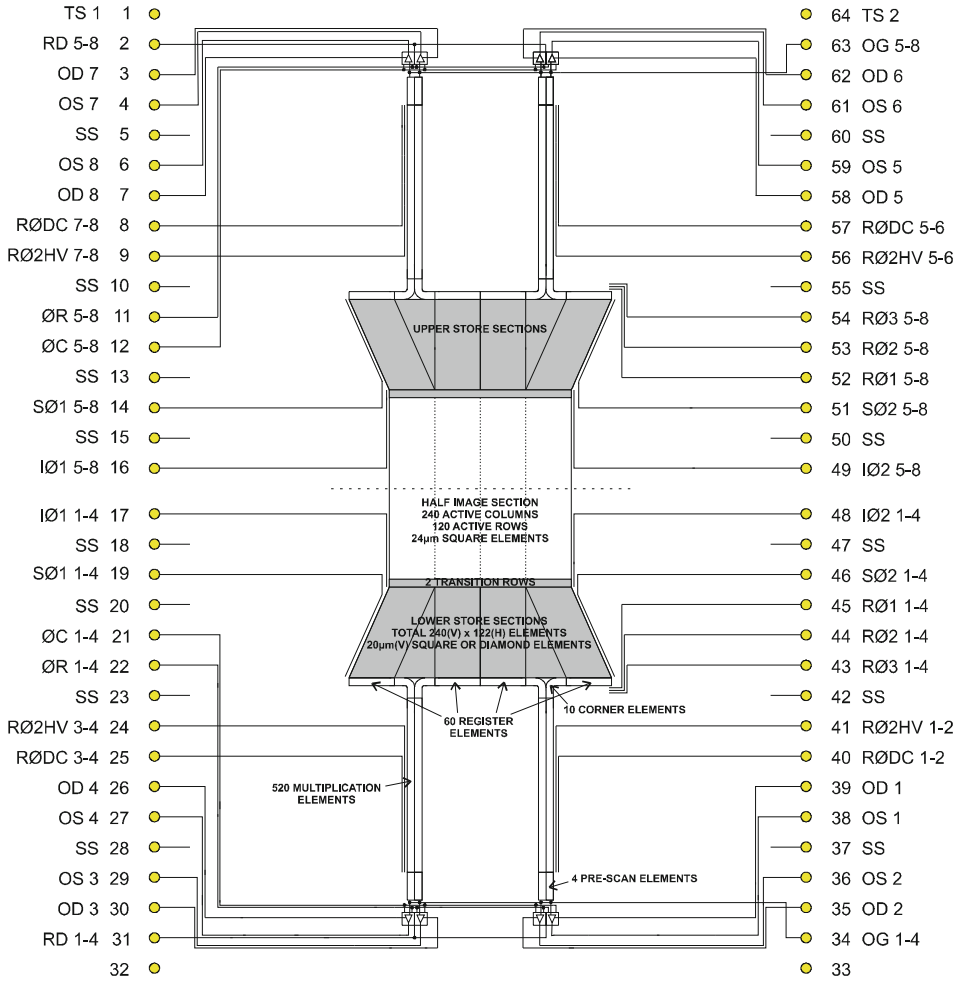
Astronomical wavefront sensors [WFS] are used for active and adaptive optics systems. Particularly for the latter, frame rates are high (up to 1,000 frames per second), and signal levels are low. Large telescopes have a higher number of subapertures and therefore require relatively large sensor formats which prove challenging to build and achieve the high frame rate with low read noise.  [Figure 13-16](#) shows an example of a WFS device, the e2v CCD220. This device has a large format, which demands high pixel rates and more outputs in order to achieve appropriate frame rates. The use of higher pixel rates increases the pixel noise, and so this device uses EMCCD technology to overcome this limitation.

It should be noted that even larger formats are required for the next generation of extremely large telescopes [ELT]. A high frame rate, large format CCD would require an excessive number of outputs and also dissipate a high power which makes it unrealistic for ELT use. An answer to this problem lies in the use of an active pixel sensor [APS] as described by Dierickx et al. (2011).

5 Summary and Future Trends

5.1 Device Size

The physical size of sensors has been illustrated above. For CCDs, the largest commercial wafer size that is used in the specialist imaging industry is 6 in. (150 mm). Monolithic devices are therefore limited to sizes that can fit within this diameter. In practice, the device needs to be



■ Fig. 13-16
240 × 240 format WFS CCD

comfortably away from the edge for security of manufacture, handling, and for minimal edge effects (including those of back-thinned wafers). This means that the maximum size is less than 95×95 mm square or 120×50 mm. APS sensors on the other hand can be made on larger wafers, e.g., 8 in. (200 mm). However, options for back-thinning of the larger wafers could then be limited. In practice for either type of silicon sensor, an important consideration is the operational yield. The probability of a fatal defect (DC failure) is an exponential function of device area. High-quality fabrication plants are able to make whole wafer devices with a viable yield, although the unit device price is high.

5.2 Pixel Size

For astronomical imaging CCD, the most common CCD pixel size is 15×15 μm , with others in the 10–25 μm range also in use. Larger pixels can certainly be designed and made; they have the advantage of high signal (charge) capacity although low noise output circuits cannot handle the larger signals. Smaller pixels are limited mainly by the capability of the fabrication plant to project and process small features. Typical pixels may have three or four electrodes, and this means that submicron feature sizes need to be used. Pixels down to about 8 μm in size are considered state-of-the-art at present.

Manufacturing technology can allow smaller features but the cost of implementation is high. Since the specialist imaging and astronomy market is finite in size, it means that investment in new equipment limits the more advanced capabilities. Again, CMOS devices can be manufactured with extremely small pixels (1 μm pixel cell-phone cameras), but specialist scientific imager fabrication facilities currently use 0.18 μm features which leads to pixels down to about 5 μm in size. In the case of APS devices, the pixel size depends on how much functionality is desired in each pixel.

5.3 Readout Noise

CCD readout noise is typically $2 e^-$ rms for a high-performance sensor at low frequencies. The EMCCD type allows subelectron readout noise to be achieved and is often used at signal levels below $1,000 e^-$. Both technologies are well established with only modest improvements likely. APS devices are being developed rapidly, and subelectron read noise is conceivable. However, attaining such a high performance probably comes at the expense of more sensitivity to pixel-pixel performance variation.

5.4 Spectral Response

As discussed above, almost 100% QE is already achieved. Advances in AR coating are allowing the breadth of spectral response to be increased, allowing higher efficiency optimized instrument designs.

5.5 Associated Electronics

As device sizes increase, the number of readout channels increase and put more demands on electronics performance. For example, large mosaics of multichannel sensors are being made, and achieving detector-limited performance for all channels is nontrivial. Local signal processing and increased digital processing are useful.

Acknowledgments

e2v is acknowledged for many illustrations used here and multiple constructive inputs from colleagues.

References

- Dierickx, B., et al. 2011, International image sensor workshop proceedings, Hokkaido e2v. www.e2v.com
- Holland, S. E., et al. 2007, Nucl Instrum Method Phys Res A, 579, 653–657
- Janesick, J. R. 2001, Scientific Charge-Coupled Devices (SPIE, Bellingham)
- Jorden, P. R., et al. 2010, Proc SPIE, 7742–19
- LSST. <http://www.lsst.org/lsst/gallery/camera>
- Mackay, C. D. 2012, Proc SPIE, 8453–01
- McLean, I. S. 2008, Electronic Imaging in Astronomy: Detectors and Instrumentation, (2nd ed.; Springer, Heidelberg)
- Pan-starrs. <http://pan-starrs.ifa.hawaii.edu/public/design-features/cameras.html>
- Stern, R. A., et al. 2004, Proc SPIE, 5171, 77
- Tulloch, S. M., & Dhillon, V. S. 2010, MNRAS
- Turner, et al. 2001, A&A, 365, L27

14 Long-Wavelength Infrared Detectors

Erick T. Young

Director, SOFIA Science Mission Operations, NASA Ames
Research Center, Moffett Field, CA, USA

1	<i>Introduction</i>	566
2	<i>Figures of Merit</i>	567
3	<i>Thermal Detectors</i>	567
4	<i>Photon Detectors</i>	571
4.1	Photoconductors	571
4.2	Impurity Band Conduction Detectors	574
4.3	Readouts	577
4.4	Detector Arrays	578
5	<i>Microwave Kinetic Induction Detectors</i>	582
6	<i>Summary</i>	584
	<i>References</i>	584

1 Introduction

Long wavelength infrared, defined in this review as the range from 5 to 1,000 μm , has proven to be essential in our understanding of many key topics and processes in astronomy. While our understanding of the universe has been limited to visual wavelengths for most of human history, the development of infrared technology has greatly expanded the study of planetary atmospheres, the interstellar medium, star formation, stellar evolution, galaxies, and the distant universe.

With eight octaves of wavelength coverage in the long wavelength infrared, it is not surprising that a wide range of different physical phenomena have been employed to detect the radiation. The topic of infrared detectors is large, and a number of excellent reviews of the technology are available. In particular, Rieke (2003, 2007), Amico et al. (2004), and McLean (2008) provide comprehensive coverage of recent developments.

Infrared detector technology is a particularly fast changing topic, and the traditional astronomical journals are generally not the first place to find papers on the latest developments in the field. The Proceedings of the SPIE which document papers presented at their topical symposia have become an important source of current information. The symposia combine both the astronomical and engineering disciplines and allow interaction of both communities. As Rieke (2009) has pointed out, much of the early development of infrared astronomy was led by experimental physicists and engineers, and that tradition continues to an extent to this day. In fact, much of the development in infrared detectors has been driven by fields other than astronomy, notably military and industrial applications. The notable exception, however, has been far-infrared and sub-millimeter detectors. At the longest wavelengths, there have been few applications other than astronomy, and the key technological advances have been motivated by astronomy.

The advancement of detector technology is often associated with funding from particular projects. For astronomy, the NASA and ESA space infrared missions have driven the field to new levels. While the basic technology usually exists when the projects are formulated, the specific performance demands of astronomy have usually required significant additional development to advance the state of the art to the required levels. Key examples have been the operation of low-background photoconductors for IRAS, the development of large format far infrared arrays for Spitzer, and the development of very high-performance arrays for the James Webb Space Telescope.

The evolution of long wavelength detector technology has followed two complementary themes. First, the inherent sensitivity of detectors has improved dramatically. Thermal infrared astronomy emerged as an essential technique with the development of the germanium bolometer in the early 1960s by Frank Low (Low 1961). At that point, the key sensitivity measure, the Noise Equivalent Power or NEP, was in the range $10^{-12} \text{ W Hz}^{-1/2}$. The current state of the art is nearly eight orders of magnitude more sensitive. The second theme is the development of multi-pixel arrays. While early infrared astronomers struggled with single pixel maps, megapixel arrays in the thermal infrared are now available. Array technology has been a direct beneficiary of the advancements in the semiconductor industry. Although arrays of highly sensitive detectors may seem commonplace to the generation reading this review, the dual developments have had in fact a truly revolutionary effect on the field.

2 Figures of Merit

Depending on the detector type, infrared detectors are subject to different noise sources. Noise can arise from the detector dark current, phonon noise in thermal detectors, amplifier noise, and ultimately from the background fluctuations. The goal of all astronomical detector systems is to be limited by the backgrounds associated with the measurement in question. For ground-based observations, the huge background associated with the warm telescope and atmosphere will dominate. In space, the ultimate background limits will be set by scattered sunlight, zodiacal thermal emission, emission from interstellar dust, or the far infrared background depending on wavelength and direction of observation.

The most commonly used figure of merit in the thermal infrared is the Noise Equivalent Power (NEP). The NEP is defined as the infrared power that yields a signal to noise ratio in the sensor of unity in 1 Hz of bandwidth, which is equivalent to $1/2$ second of integration time.

Another key figure of merit is the detective quantum efficiency η_d . It compares the NEP of the detector in question with the NEP of an idealized detector only limited by the fluctuations in the background.

$$\eta_d = (NEP_{BLIP}/NEP)^2$$


where NEP_{BLIP} is the NEP of the ideal background-limited infrared photodetector and is given by:

$$NEP_{BLIP} = (4P_{BG}hc/\lambda)^{1/2}$$

where P_{bg} is the background power on the detector, h is Planck's constant, c is the speed of light, and λ is the average wavelength. In the absence of other noise sources, the NEP of a detector in fact scales as the inverse square root of the efficiency with which it converts photons to electrical signals, hence the quantum efficiency name.

3 Thermal Detectors

The first infrared detectors were thermal detectors, where the infrared radiation produces a change in the temperature of the detector, and this temperature change is measured. Among the early thermal detectors were the thermopile (Coblentz 1914) and the vacuum thermocouple (Pettit and Nicholson 1922). The first important astronomical measurements in the infrared were done with thermal detectors. Depending on the sophistication of the temperature sensor, the sensitivity can be very low (for the case of Herschel's thermometer) to exquisitely high (for modern superconducting bolometers).

The most important form of thermal sensor is the bolometer. All bolometers share common characteristics.  Figure 14-1 illustrates the key elements. Following the simplified bolometer theory first formulated by Low (1961), there is an absorber of heat capacity C that intercepts the infrared radiation. It is connected to a cold sink by a link of thermal conductance G . When radiation hits the absorber, its temperature rises, and this is sensed by the thermometer. In steady state, the amount of temperature rise ΔT is simply given by P/G , where P is the absorbed power. Since the temperature rise is inversely proportional to G , utilizing very low conductance links can lead to high sensitivity. Unfortunately, the time constant of response τ is given by

$$\tau = C / G$$

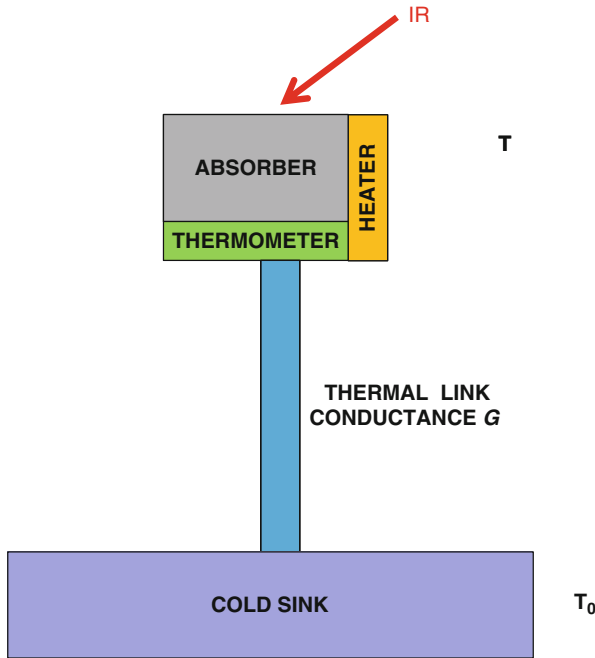


Fig. 14-1
Diagram of generic bolometer

Hence, there is a premium to making the heat capacity of the absorber as small as possible to avoid excessively slow response times.

The fundamental limit on the performance of a bolometer is from the thermal fluctuations in the thermal link between the cold sink and the absorber. Hence, all high-performance bolometer systems are cooled to very low temperatures, often into the sub-Kelvin regime. Most of the design effort in bolometers has been in producing the most sensitive thermometers, the smallest thermal links, and the smallest thermal masses of the absorber and thermometer. The key figure of merit for a bolometer is the power of the infrared signal equal to the noise of the bolometer. The Noise Equivalent Power (NEP) of a bolometer is given by

$$NEP = (4kT^2G)^{1/2}/\eta$$

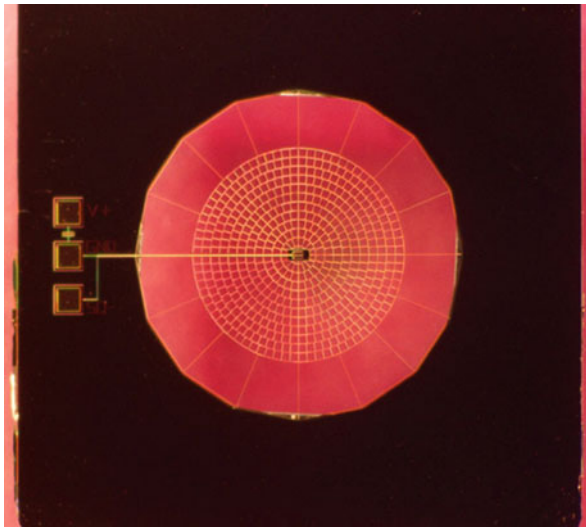
where k is Boltzmann's constant, T is the temperature, and η is the quantum efficiency. Bolometer theory has been refined in the years since Low's treatment to include more detailed modeling of the nonlinear effects of thermal gradients in G as well as the temperature dependence of the parameters (see, e.g., Mather 1982, 1984, and Richards 1994).

The desire to maximize sensitivity by making the thermal conductance G as small as possible is complicated by corresponding increase in the thermal time constant $\tau = C / G$ of the bolometer, where C is the heat capacity of the bolometer element. An important improvement in the response speed of bolometers was realized by making the bolometer part of a thermal feedback circuit, as illustrated in the figure. The bolometric element is heated slightly above the bath temperature using a heater. Any required temperature changes due to the absorption of

infrared radiation are sensed by the thermometer, and a corresponding reduction in the heater power is applied by the feedback circuit. This *electrothermal feedback* reduces the effect of the heat capacity of the bolometer since the temperature is no longer required to change, and the effective time constant is reduced by the loop gain of the feedback circuit.

Various thermometer elements have been used in astronomical bolometers. The first highly sensitive bolometers were developed by Low (1961) and utilized gallium-doped germanium as the sensor. When cooled to liquid helium temperatures, this material exhibits large changes in electrical conductivity with temperature and was a mainstay of infrared photometers for many years. Suitable thermometric elements have been produced with both direct doping of germanium and neutron transmutation doping of pure germanium to produce the desired doping levels (Haller et al. 1994). More recently, doped silicon bolometers have been used on a number of applications.

At longer wavelengths, the absorbing mass of the bolometer can be significantly reduced by using the spiderweb design. ● Figure 14-2 shows a bolometer that was designed for use on the Herschel Spectral and Photometric Imaging REceiver (SPIRE) instrument. The absorber is a dielectric film that has been coated with a thin absorbing metallic layer. As shown in the figure, most of the film has been etched away, leaving only a suspended web. The neutron transmutation doped germanium thermistor is located at the center of the web. This design has a number of significant advantages over previous architectures, particularly for a space application. First, the reduced mass results in high sensitivity. Second, the resonant mechanical frequency of the structure is much higher than in a conventional design, an important consideration in the launch environment. Finally, because the spiderweb bolometer is primarily empty space, the





■ Fig. 14-2

Spider web bolometer designed for Herschel SPIRE instrument. The absorbing element is the suspended mesh, and the thermometer element is a neutron transmutation doped germanium chip in the center (J. Bock, JPL)

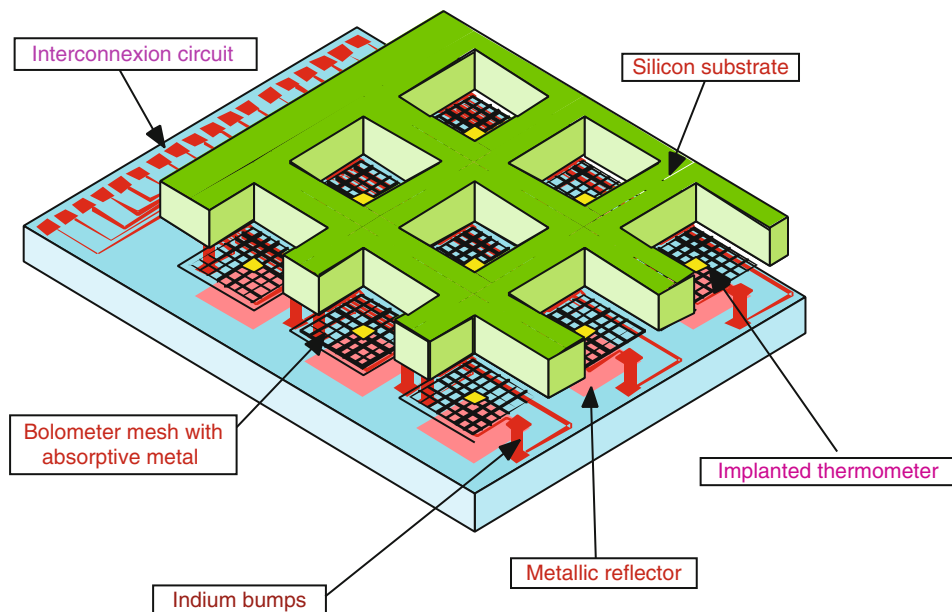
cross section for interaction with cosmic rays is greatly reduced. Spiderweb bolometers have been flown on both the Herschel SPIRE instrument (Griffin et al. 2010) and High Frequency Instrument (HFI) on the Planck mission (Holmes et al. 2008; Planck et al. 2011).

The semiconductor thermometers are typically biased at a constant current, and voltage changes associated with temperature changes are detected. For systems of this type, the signal levels are usually very low, and J-FET amplifiers have been the only devices with low enough noise to be useful in the most demanding applications. Since J-FETs cannot operate at the same sub-Kelvin temperatures required by the detectors, significant engineering is required to provide both thermal and optical isolation for the bolometer (running at <300 mK) from the associated amplifier (running at ~ 100 K). These complications have generally limited the application of semiconductor bolometer to modest sized arrays, typified by the 384-element SHARC-II and HAWC arrays (Moseley et al. 2004).

A much larger semiconductor bolometer array was used on the PACS instrument (Poglitsch et al. 2010) on the Herschel Space Observatory. The imaging photometer section of PACS has two large bolometer arrays of 32×16 pixels and 64×32 pixels made up of 16×16 sub-modules. Their design has been described in Agnese et al. (2003) and Billot et al. (2007, 2009). The entire bolometer array is fabricated out of silicon that is micro-machined to form the necessary structures. The bolometer design is based on a two-dimensional screen situated $\lambda/4$ above a metallic backshort. The standing wave pattern with the backshort results in a maximum amplitude in the electric field at the screen. The screen is coated with an absorber that matches the impedance of free space, resulting in highly efficient absorption of the energy. The thermometer for the detector consists of ion-implanted silicon semiconductor. The other key difference in the PACS bolometer array is the use of MOSFET amplifiers operating at the detector temperature. Normally, the voltage noise of MOSFETs is too high to be useful, but this noise is overcome by making the responsivity of the very high by making the resistance of the ion-implanted thermometer very high, in the 1–10 G Ω range (Agnese et al. 2002).  Figure 14-3 diagrams the key elements of the PACS photometer array (Agnese et al. 2002).

An important development in astronomical bolometer detectors has been the use of superconducting films to sense the temperature changes caused by photons. These systems take advantage of the extremely steep dependence of the film resistance with temperature near the superconducting transition. By sensing the changes in current flow with changes in illumination, these transition edge sensor (TES) bolometers make exceedingly sensitive infrared detectors.  Figure 14-4 is a functional diagram of a TES bolometer. A superconducting film that is deposited on the thermal link is the key part of a voltage-biased sensing circuit. The equilibrium condition has part of the film being warm enough to be a normal conductor and part of the film superconductive. As a normal conductor, current in that part of the circuit will dissipate heat, keeping it normal. Input infrared radiation will try to raise the temperature of the absorber, forcing more of the film to go normal. Since an increase in the amount of normal film increases the electrical resistance in the circuit, the power dissipation in the film drops to just compensate for the infrared radiation. Thus, properly designed TES bolometers inherently have electrothermal feedback.

TES detectors lend themselves naturally to the use of Superconducting QUantum Interference Devices (SQUIDS) as the current measuring device in the circuit (e.g., Irwin et al. 2004). As superconducting devices that operate at the same temperature regime as the TES bolometers, thermal design is straightforward. Recent examples of TES detectors are the Goddard IRAM Superconducting 2-mm Observer (GISMO) (Staguhn et al. 2008) and the Submillimetre Common-User Bolometer Array (SCUBA-2) instrument (Bintley et al. 2010).



■ Fig. 14-3

Diagram of PACS photometer bolometer array concept. The etched, metalized silicon mesh sits $1/4$ wavelength above the reflective backshort. The ion-implanted silicon thermometer (yellow) is connected to the readout electronics through indium bump bonds (Agnese et al. 2002)

4 Photon Detectors

4.1 Photoconductors

Photoconductors are semiconducting materials that produce free charge carriers when photons of sufficient energy hit the material. ☉ *Figure 14-5* illustrates the simplest photoconductor circuit. An electrical bias is applied to a photoconductor, and the current is measured by some electronic means.

The minimum energy to generate free charge carriers is known as the band gap E_g of the semiconductor and corresponds to the energy between the conduction and valence bands of the material. The longest wavelength that can be detected by a photoconductor is related to the band gap by:

$$\lambda_c = hc/E_g = 1.24(\mu\text{m})/E_g(\text{eV})$$

where h = Planck's constant and c = speed of light. For silicon, the band gap is 1.1 eV, meaning that the longest infrared wavelength that can be detected is just over 1.1 μm . Fortunately, other semiconductor materials have smaller band gaps. The most important materials for infrared astronomy have been indium antimonide with a 5.3 μm cutoff and mercury cadmium telluride. For mercury cadmium telluride, the band gap can be adjusted by varying the fraction x of the cadmium vs. mercury in the material ($\text{Hg}_{(1-x)}\text{Cd}_x\text{Te}$). In this way, high-performance detectors have been built with cutoff wavelengths from as short as 1 μm to beyond 10 μm .

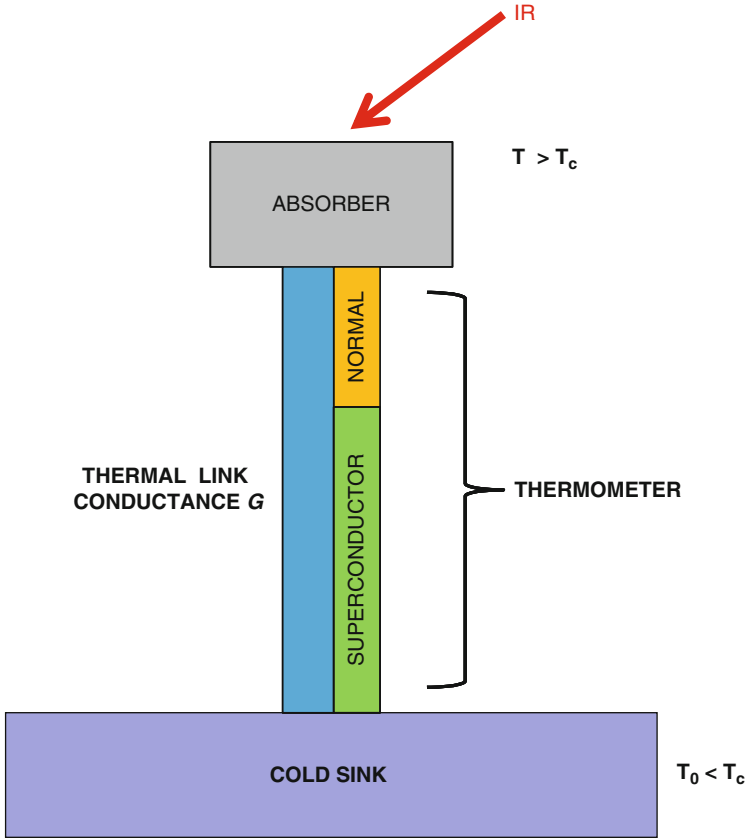


Fig. 14-4
Functional diagram of a transition edge superconducting bolometer

Table 14-1
Some important intrinsic infrared detector materials

Intrinsic photoconductors	
Material	Cutoff wavelength (μm)
HgCdTe	0.8 to >20
Si	1.1
Ge	1.6
InSb	5.5

In addition to utilizing the intrinsic band gap of semiconductor materials, detectors have employed the energy levels associated with dopants in semiconductors. Figure 14-6 shows a schematic band gap diagram for silicon doped with arsenic. The impurity energy levels of arsenic are only 0.054 eV from the conduction band, meaning that a photon of energy greater than 0.054 eV (wavelength less than 23 μm) would be energetic enough to ionize a charge carrier into the conduction band. This process is called extrinsic photoconductivity. Important

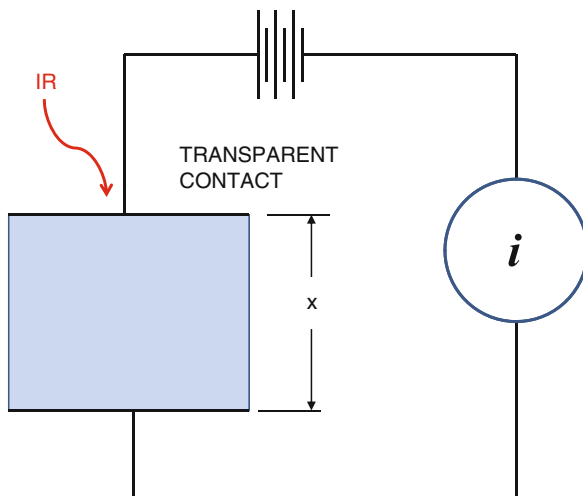


Fig. 14-5
Simple photoconductor circuit

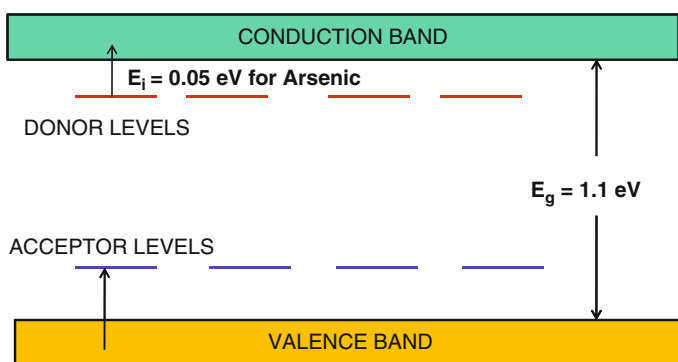


Fig. 14-6
Band gap diagram of silicon showing extrinsic impurity levels of arsenic

reviews of extrinsic photoconductors can be found in the papers by Bratt (1977) and Sclar (1984). Table 14-2 lists some of the important extrinsic dopants that have been used in infrared astronomy.

The photoionization cross section $\sigma(\lambda)$ is a measure of the probability that an impurity atom is ionized by an incoming photon at a given wavelength λ . For an impurity of concentration N atoms per cm^3 , the absorption coefficient $\alpha(\lambda)$ just $\alpha(\lambda) = N_o \sigma(\lambda)$. The fraction of radiation that is absorbed in a path length x is then $1 - e^{-\alpha x}$. Referring back to Figure 14-5, if R_1 is the reflectivity of the first surface of the detector and R_2 is the reflectivity of the back surface, the probability of a photon being absorbed in the detector is (Young 1983)

$$\eta = \frac{(1 - R_1)(1 - e^{-\alpha x})(1 + R_2 e^{-\alpha x})}{(1 + R_1 R_2 e^{-2\alpha x})}$$

■ Table 14-2

Important extrinsic dopants in silicon and germanium (Data from Bratt 1977)

Extrinsic photoconductors				
Material	Dopant	Type	Cutoff wavelength (μm)	Photoionization cross section (cm^2)
Si	Ga	p	17	5×10^{-16}
Si	As	n	23	2.2×10^{-15}
Si	B	p	28	1.4×10^{-15}
Si	Sb	n	29	6.2×10^{-15}
Ge	Be	p	52	
Ge	Ga	p	115	1.0×10^{-14}
Ge	Sb	n	129	1.6×10^{-14}

Ideally, it is highly desirable to have as high an absorption coefficient as possible, to maximize the absorption of photons in the detector. From Table 14-2, the typical photoionization cross sections are 10^{-14} cm^2 in germanium and $\sim 10^{-15} \text{ cm}^2$ in silicon. The most straightforward way of increasing the absorption coefficient is to increase the doping concentration of the extrinsic species. Unfortunately, there is a limit to the doping concentration because of the formation of an impurity band in the semiconductor. When this happens, the detector will have unacceptably high dark currents. For gallium-doped germanium detectors, the maximum doping is $2 \times 10^{14} \text{ cm}^{-3}$ (Wang et al. 1986), which means that germanium photoconductors must have absorption path lengths of several millimeters for reasonable efficiencies.

An important variation of the bulk germanium photoconductor is the stressed detector. The long wavelength response of Ge:Ga can be extended by applying uniaxial stress along the (100) axis (Kazanskii et al. 1977). This technique has been used on both the MIPS instrument on Spitzer and the PACS instrument on Herschel. With the application of 490 MPa of stress, the MIPS stressed detectors have useful response out to 200 μm (Young 2000). The design of a stressed detector focal plane presents a challenging combination of mechanical, thermal, optical, and electronic design. The methodology used for the PACS instrument has been described by Rosenthal et al. (2002).


Bulk photoconductors are subject to a number of significant nonlinearities, particularly at low backgrounds. The fundamental reason is that at the very low backgrounds found in many astronomical applications (particularly space astronomy), the detectors are essentially insulators, and the time constants associated with reaching electrical equilibrium are large compared with typical measurement times. These effects have been Bratt (1977) and Sclar (1984). In addition, bulk photoconductors are subject to large response changes when exposed to ionizing radiation. Significant effort has been devoted to characterizing and calibrating these effects for space missions. For the Spitzer MIPS arrays, the calibration strategy included operating in well-defined time domains and frequent use of internal calibration sources (Gordon et al. 2005, 2007; Stansberry et al. 2007).

4.2 Impurity Band Conduction Detectors


While in principle it is possible to make sensitive infrared detectors using simple photoconductors, this mode of operation is almost never used. Dark current due to impurities in the

material will usually limit the performance of these devices. Practical modern direct detectors use more complicated structures. For the intrinsic photoconductors, the detectors are almost always made as photodiodes, where n-doped material forms a junction with p-doped material. By forming a p-n junction, a charge-free region, the depletion region, is formed, and any photoexcited charge carriers formed in this region are swept to the contacts by the electric field in the region. High-performance photodiodes have been formed in HgCdTe for wavelengths as long as 10 μm (Bacon et al. 2004). Photodiodes are the most important type of near-infrared detector, and they are covered by the chapter in this volume by Beletic.

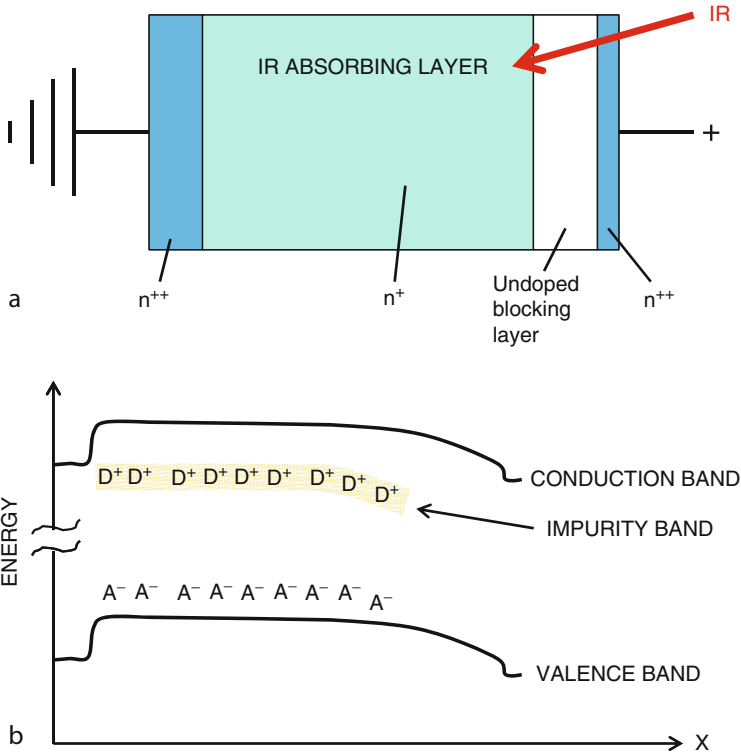
From $\sim 10 \mu\text{m}$ to nearly 40 μm , the most sensitive detectors have used the impurity band conduction (IBC) structure. Also known as blocked impurity band (BIB) detectors, these devices overcome many of the fundamental limitations of classical bulk photoconductors. The photon absorption efficiency of a photoconductor is set by the product of the photoionization cross section of the dopant and the doping concentration of the desired impurity. The latter is limited by excessively high dark currents when the concentration gets too high.

In IBC detectors, the infrared absorbing layer is doped well beyond the point where an impurity band forms but not so high that degenerated metallic conduction occurs. In fact, dopings as high as $100 \times$ greater than possible for bulk photoconductors are typically employed. Thus, detector structures can be $100 \times$ thinner for comparable absorption efficiencies. To block the very high dark current, IBC detectors incorporate a very pure undoped “blocking” layer in series with the doped layer. Since the free carriers in the impurity band cannot reach the contact, the dark current is blocked. Only the carriers that are excited into the conduction band (for an n-type IBC detector) reach the external circuit.  Figure 14-7 schematically shows band diagram of an IBC detector. With bias, negative charge carriers are injected at the N^{++} contact and migrate toward the positive contact via hopping conduction until they reach the undoped layer. Since there is no impurity band in this layer, the negative carriers are blocked and accumulate at the interface. This accumulation of charge creates a local field that clears out free charge carriers in the heavily doped region. For photon-excited carriers in the heavily doped region to reach the contacts, they must be driven by the electric field in the region, and hence this depletion region is necessary for the operation of the device. This region has a width “ w ” and is given by (Petroff and Stapelbroek 1984, 1986)

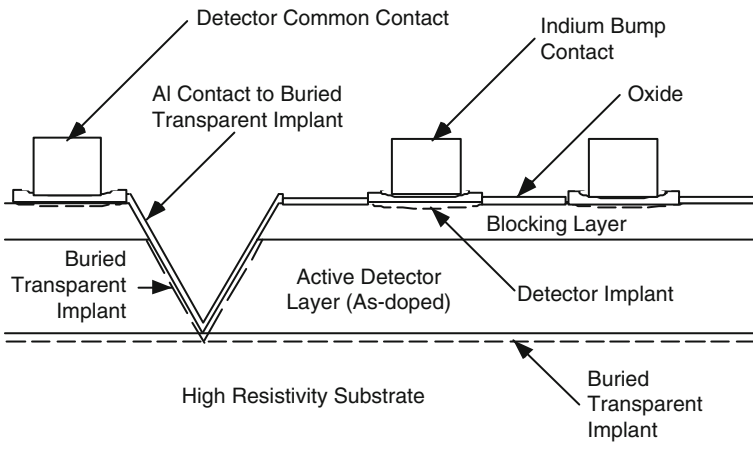
$$w = \left[\frac{2\epsilon\epsilon_0(V_a - V_{bi})}{eN_A} \right]^{1/2} - b$$

where ϵ is the dielectric constant, ϵ_0 is the permittivity of free space, e is the electron charge, V_a is the applied bias voltage, V_{bi} is the built-in potential of the undoped active layer junction, and b is the blocking layer thickness. The width of the depletion region determines how much of the heavily doped region usefully converts photons to detected charge carriers. For this depletion region to be successfully created, the background impurity level N_A of acceptors (A in the figure) must be exceedingly low. Fortunately, the technology of silicon epitaxial growth is up to the task. An example of the architecture of a modern IBC detector is shown in  Fig. 14-8. A detailed analysis of the IBC detector has been given by Szmulowicz and Madarsz (1987).

An additional advantage to the high doping in the IR absorption layer is an extension of the cutoff wavelength beyond the normal extrinsic limit. The finite energy width of the impurity band means that there are energy states closer to the conduction band than in a normal photoconductor. For Si:As IBC detectors, the cutoff is extended to $\sim 28 \mu\text{m}$, while for Si:Sb IBC detectors, the cutoff is $\sim 36 \mu\text{m}$.



■ Fig. 14-7
 (a) Representative structure of n-type IBC detector, (b) band diagram of the detector (Haller and Beeman 2002)



■ Fig. 14-8
 IBC detector cross section (Love et al. 2010)

4.3 Readouts

A key part of the detector system is the method for measuring the photocurrent. A number of different circuits have been used, and a sample are illustrated in [Fig. 14-9](#).

One of the earliest circuits was the transimpedance amplifier (TIA). Here, the photocurrent nulled at the input by a corresponding opposite current that is driven by the voltage across a resistor in a feedback loop. One advantage of this circuit is that because of the negative feedback, the bias potential across the detector is constant. The output signal is simply

$$V_{OUT} = i_d R_f$$

where i_d is the photocurrent and R_f is the feedback resistance. The main limitation of the TIA is that the feedback resistor produces Johnson noise

$$V_{n,J} = (4k_B T R_f)^{1/2}$$

where k_B is Boltzmann's constant and T is the temperature of the resistor. The TIA was used on the IRAS focal plane array.

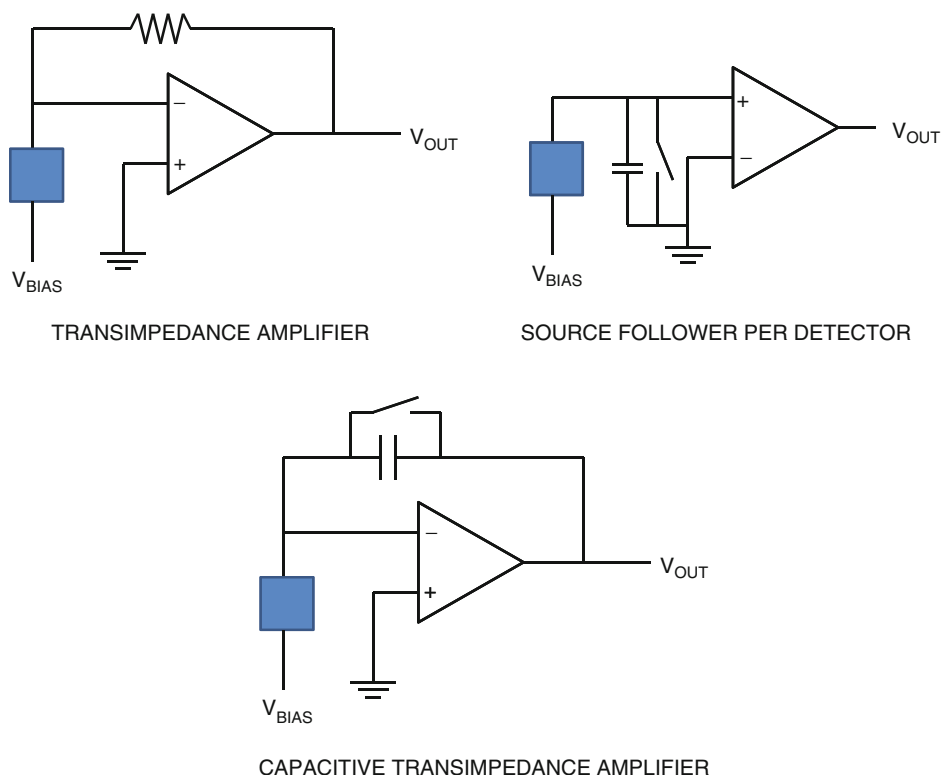


Fig. 14-9

Some representative readout circuits used with photoconductors

A variant of the TIA substitutes a capacitor for resistor. In the capacitive transimpedance amplifier (CTIA), the output signal is given by:

$$V_{\text{OUT}} = Q/C_f$$

where Q is the total charge collected on the feedback capacitor C_f . Using modern semiconductor techniques, very small feedback capacitors are possible, making such circuits potentially very sensitive. Since the charge on the feedback capacitor cannot be built up indefinitely, the capacitor needs to be occasionally reset using a reset switch. The principal drawback to the more widespread application of the CTIA is the need to build a high gain, continuously operating inverting amplifier for each pixel with its attendant real estate and power dissipation. Despite these drawbacks, the CTIA has been successfully used in high-performance astronomical applications. Most notably, the CTIA was used as the readout for the MIPS 70 and 160 μm arrays on Spitzer. For both applications, the constancy of bias voltage for the germanium photoconductors was an overriding consideration.

The need for simplicity and low power dissipation, particularly as array formats have increased, has led to the widespread adoption of the source follower per detector (SFD) circuit. Here, the voltage is developed across a capacitor that is occasionally reset. The output is

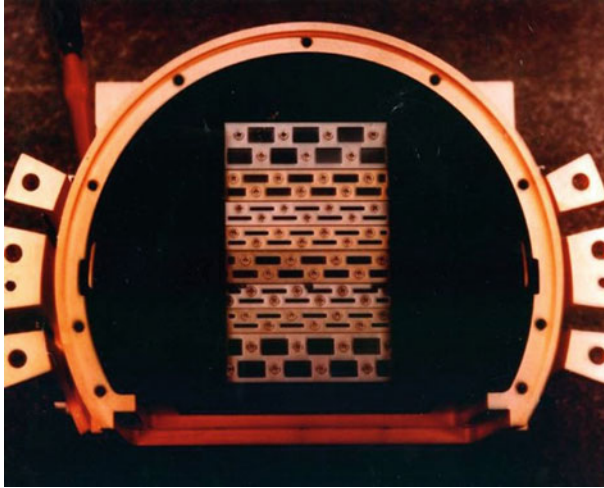
$$V_{\text{OUT}} = AQ/C_{in}$$

where A is the voltage gain of the circuit, typically of order unity. The input capacitor C_{in} now includes the amplifier input capacitance as well as the detector capacitance. The SFD can be implemented in very few transistors and has become the most common circuit in very large arrays. An important power saving characteristic of the SFD is that the amplifier that measures the voltage on the capacitor only needs to be turned on for the brief time it takes to make the measurement. For a large array, the effective duty cycle can then be exceedingly low. The main disadvantage of the SFD is that as charge accumulates on the input capacitor, the voltage across the detector changes, leading to inevitable nonlinearities. For most detectors, the charge accumulation is limited so the total voltage change is only a fraction of the applied bias.

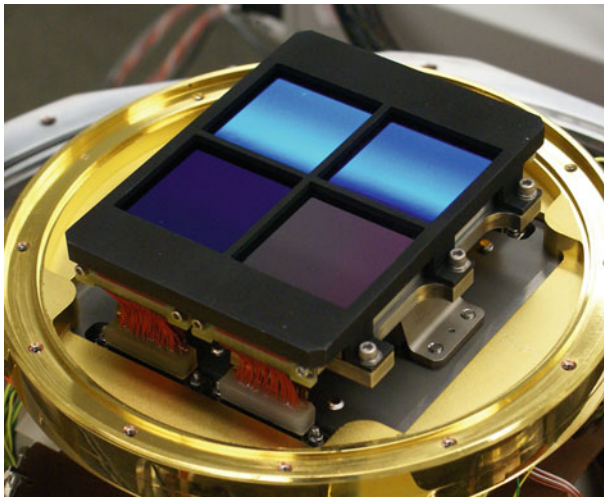
4.4 Detector Arrays

Arguably, one of the most important developments in infrared astronomy has been the move from single detectors to arrays. Increases in overall information gathering power of many orders of magnitude have been realized over the past few decades. Very much following the advances semiconductor fabrication technology, infrared detector arrays have seen an exponential growth in array format mimicking Moore's Law (Moore 1965). A good illustration of the advances in just three decades is to compare the 63-detector IRAS focal plane array (► Fig. 14-10) with one of the focal plane arrays which has 16 million pixels for the JWST NIRCam instrument (● Fig. 14-11). The entire NIRCam instrument has 40 million pixels.

The key to this continuing advance has been, in fact, the utilization of many of the technologies from the semiconductor industry. ● Figure 14-12 illustrates the principal architecture used on large format detector arrays. Using integrated circuit technology, a readout integrated circuit (ROIC) is fabricated out of silicon. This circuit typically has a unit cell amplifier for each pixel and all of the multiplexing circuitry needed to send the signals down a manageable number of output lines. The inputs to all these amplifiers are arranged in a grid on the top of the integrated circuit. The detector array is constructed out of the appropriate material for infrared absorption and carrier generation. In some cases, the detector is an array of photodiodes

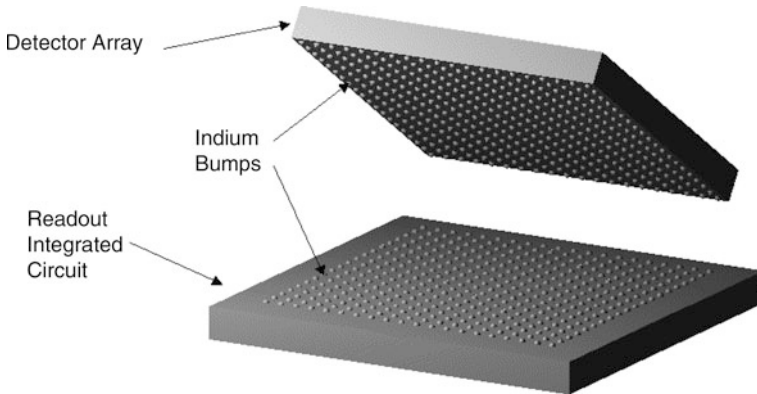


■ Fig. 14-10
IRAS focal plane array consisting of 63 discrete photoconductors



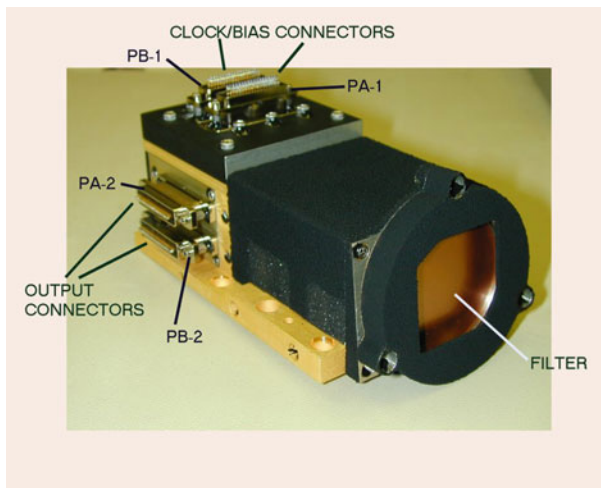
■ Fig. 14-11
One of two 1–2.5 μm focal plane arrays that will be used on the NIRCam instrument on JWST. It is a 2×2 mosaic of $2,048 \times 2,048$ pixel HgCdTe arrays. The complete instrument has a pair of these mosaics in addition to two $2,048 \times 2,048$ pixel 2–5 μm focal planes, for a total of 40 million pixels

fabricated out of HgCdTe, and in other cases, the detector could be an array of IBC detectors on a silicon substrate. The need to connect the outputs of the detectors to the inputs of the amplifiers is accomplished with indium bump bonds. These are exactly matched microscopic indium bumps that are deposited on both the detector array and the ROIC. When the two sets of bumps are aligned and pressed together, a robust cold weld is formed in the indium. This technology has reached the level of a fine art, with as many as four million interconnections being made



■ Fig. 14-12

Indium bump bonding of an infrared detector array to a silicon readout integrated circuit is the principal architecture for detectors in the 1–40 μm range

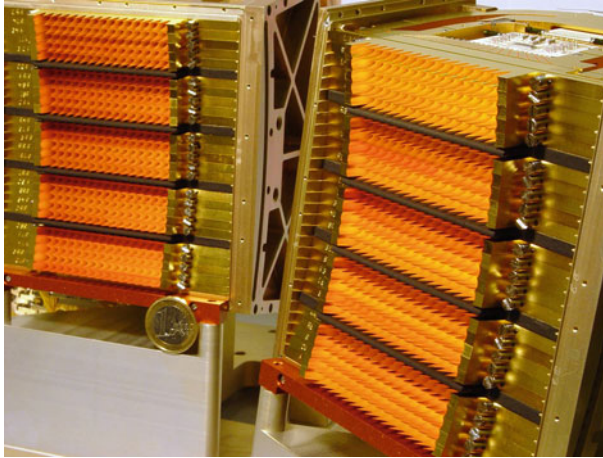


■ Fig. 14-13

Spitzer MIPS 70 μm focal plane array (E. Young)

with very high success rate (>99%) and high reliability. This simplistic description, of course, bypasses many of the difficult engineering development that has been needed to make the whole process work. In particular, low-temperature operation of the readout, accommodation of the thermal properties of materials, and the development of good infrared detector materials have been the work of many scientists and engineers.

At wavelengths beyond the silicon detector range ($\lambda > \sim 40 \mu\text{m}$), the road to larger arrays has been slower. To date, despite some promising early work (Watson and Huffman 1988; Rossington 1988), the IBC technology has not been successfully extended to germanium detectors. Instead, bulk photoconductors have been used on Spitzer MIPS instrument and the PACS



■ Fig. 14-14
Photoconductor arrays for the PACS instrument (ESA)

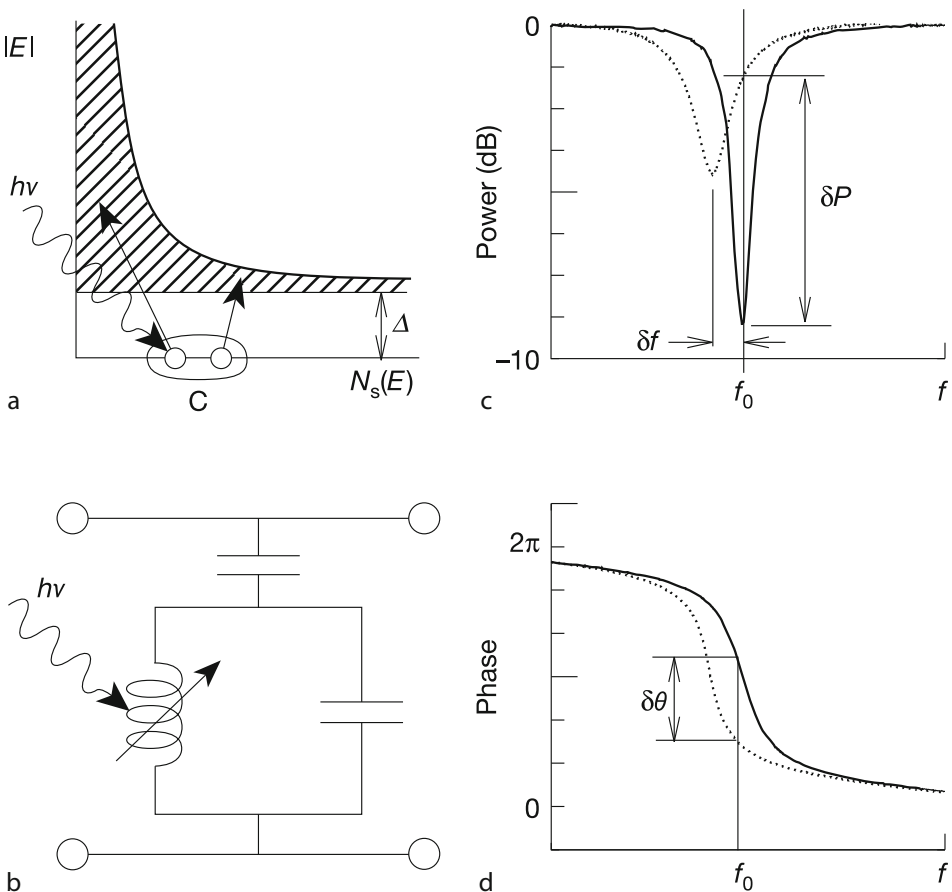
■ Table 14-3
Some representative long wavelength detector arrays

Array	Material	Format	Wavelength (μm)	Operating temp (K)	Notable example	Ref
Teledyne H2RG (JWST)	HgCdTe	$2,048 \times 2,048$	0.6–5.3	38	JWST NIRSPEC and NIRCAM	Smith et al. 2009
Teledyne H4RG	HgCdTe	$4,096 \times 4,096$	0.6–5.3	38		Blank et al. 2011
Raytheon orion	InSb	$2,048 \times 2,048$	0.6–5.5	32		Hoffman et al. 2004
DRS MEGAMIR	Si:As Si:Sb	$1,024 \times 1,024$	5–28, 5–38	7	SOFIA FORCAST	Mainzer et al. 2005
Raytheon JWST MIRI	Si:As	$1,024 \times 1,024$	5–28	7	JWST MIRI	Ressler et al. 2010
CEA Si bolometer	Si	$32 \times 16,64 \times 32$	60–85, 85–130, 130–210	0.3	Herschel PACS	Billot et al. 2010
SCUBA-2 bolometer	TES	$4 \times 32 \times 40$	850, 450	0.3	SCUBA-2	Bintley et al. 2010

spectrometer on Herschel. ● *Figure 14-13* shows the $70 \mu\text{m}$ 32×32 pixel array, and ● *Fig. 14-14* shows the stressed photoconductor arrays for PACS. For both developments, custom CTIA readouts functional at 1.5 K were required. Some representative detector arrays are given in ● *Table 14-3*.

5 Microwave Kinetic Induction Detectors

An active area of detector development has opened with the advent of the microwave kinetic induction detector (MKID). Superconductors below their critical temperature T_c carry super-current by pairs of electrons known as Cooper pairs. The interaction energy 2Δ of this Cooper pair is roughly $k_B T_c$, where k_B is Boltzmann's constant and Δ is the superconducting energy gap. • [Figure 14-15](#) illustrates the operating principles of the MKID detector. While a superconductor has a zero D.C. impedance, the Cooper pairs in the superconductor present a surface inductance to A.C. signals due to the inertia of the Cooper pairs. The breaking of the Cooper pairs by the infrared photons of energy greater than 2Δ creates unpaired electrons or quasiparticles resulting in a change in the kinetic inductance of the film. Microwave kinetic induction



■ Fig. 14-15

MKID operating principles. (a) Energy diagram for an superconducting film. A photon breaks the Cooper pair C into two quasiparticles. (b) Equivalent circuit for the MKID resonator. (c) Change in output of resonator with creation of quasiparticles as resonant frequency shifts from f_0 . (d) Corresponding change in phase of the output signal (Day et al. 2003)

detectors (MKIDS) utilize this creation of quasiparticles in a superconducting film to detect photons (Day et al. 2003). A comprehensive review of the physics of MKID detectors can be found in Zmuidzinas (2012).

To sense these quasiparticles, the superconductor is made part of a highly tuned resonant microwave circuit that is excited by an external oscillator, typically in the 1–10 GHz range. The creation of quasiparticles causes an increase in surface inductance, shifting the resonant frequency lower. The change in inductance can be sensed by observing either amplitude or phase changes in the output. Because resonators made of superconductors can have exceedingly low losses, very high Q factors are possible (Mazin 2005).

The promise of MKID detectors is that it is possible to frequency multiplex many detectors on a single output line by having each pixel tuned to a slightly different microwave frequency. This tuning is accomplished geometrically in the fabrication using standard micro-lithographic techniques. In operation, each pixel is excited by a tone matched to its resonant frequency as illustrated in [Fig. 14-16](#). Hence, each pixel is associated with single microwave frequency that can be analyzed by a microwave spectrometer.

While [Fig. 14-16](#) shows individual oscillators for each pixel, they can instead be replaced by a single high-speed digital to analog converter that puts out a waveform with the necessary Fourier components to replicate the desired comb spectrum.

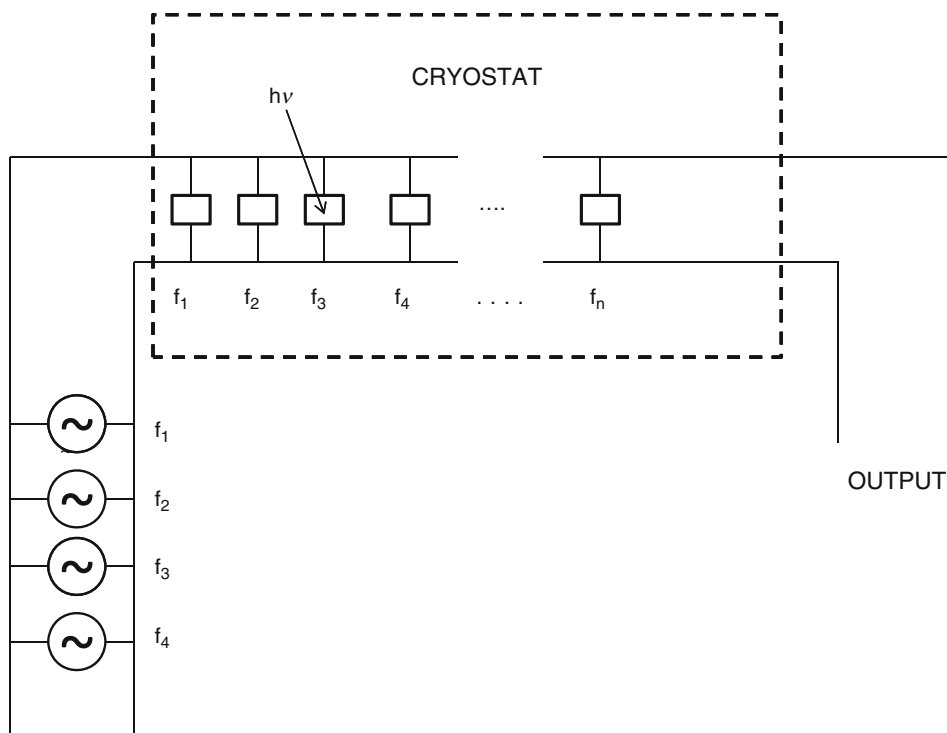


Fig. 14-16
Frequency multiplexed MKID array

MKID arrays are an active area of detector work, and a number of cameras for ground-based telescopes have been developed including the Caltech Submillimeter Telescope (CSO) (Maloney et al. 2010), APEX (Heyminck et al. 2010), and IRAM (Monfardini et al. 2011).

6 Summary

In the past 40 years, long wavelength infrared detector technology has experienced a continuous, exponential growth in array format. Arrays of 16 million pixels in the near infrared and 10^4 pixels in the far infrared have been attained. This format increase has also been matched with a nearly eight order of magnitude improvement in sensitivity. This remarkable expansion in detector performance has enabled fundamentally new insights on all areas of astronomy.

References

- Agnese, P., Rodriguez, L., & Vigroux, L. 2002, Far-IR, Sub-mm & MM Detector Technology Workshop, held 1–3 April 2002 in Monterey, CA. Organized and Sponsored by NASA/Ames & USRA/SOFIA. Online at http://www.sofia.usra.edu/det_workshop.id.66
- Agnese, P., Cigna, C., Pornin, J.-L., et al. 2003, Proc. SPIE, 4855, 108
- Amico, P., Beletic, J. W., & Beletic, J. E. 2004, Scientific Detectors for Astronomy, The Beginning of a New Era (Netherlands: Kluwer)
- Bacon, C. M., McMurtry, C. W., Pipher, J. L., et al. 2004, Proc. SPIE, 5167, 313
- Billot, N., Agnese, P., Augueres, J.-L., et al. 2007, Proc. SPIE, 6275, 62750D
- Billot, N., Rodriguez, L., Okumura, K., Sauvage, M., & Agnès, P. 2009, in EAS Publications Series, 37, Astrophysics Detector Workshop 2008, Nice, ed. P. Kern, 119–125
- Billot, N., Sauvage, M., Rodriguez, L., et al. 2010, Proc. SPIE, 7741, 774102
- Bintley, D., MacIntosh, M. J., Holland, W. S., et al. 2010, Proc. SPIE, 7741, 7741106
- Blank, R., Anglin, S., et al. 2011, ASPC, 437, 383
- Bratt, P. R., 1977 in Semiconductors and Semimetals, Vol. 12, eds. R. K. Willardson, & A. C. Beer (New York: Academic), 39–142
- Coblentz, W. W. 1914, PASP, 26, 169
- Day, P., Leduc, H., Mazin, B., Vayonakis, A., & Zmuidzinas, J. 2003, Nature, 425, 817–821
- Gordon, K. D., Engelbracht, C. W., Fadda, D., et al. 2007, PASP, 119, 1019
- Gordon, K. D., Rieke, G. H., Englebracht, C. W., et al. 2005, PASP, 117, 503
- Griffin, M. J., Abergel, A., Abreu, A., et al. 2010, A&A, 518, L3
- Haller, E. E., & Beeman, J. 2002, Far-Infrared Photoconductors: Recent Advances and Future Prospects, in Proceedings of Monterey Workshop on FIR, SubMM, and MM Detectors, NASA/CP-211408
- Haller, E. E., Itoh, K. M., Beeman, J. W., Hansen, W. L., & Ozhogin, V. I. 1994, Proc. SPIE, 2198, 630–637
- Heyminck, S., Klein, B., Guesten, R., et al. 2010, Twenty-First International Symposium on Space Terahertz Technology, held March 23–25, 2010 at Oxford University's Said Business Center and the STFC Rutherford Appleton Laboratory, Oxford, UK. National Radio Astronomy Observatory (NRAO), 262
- Hoffman, A. W., Corrales, E., Love, P. J., et al. 2004, Proc. SPIE, 5499, 59
- Holmes, W. A., Bock, J. J., Crill, B. P., et al. 2008, Appl. Opt., 47, 5996
- Irwin, K. D., Audley, M. D., Beall, J. A., et al. 2004, Nucl. Inst. Method Phys. Res. A, 520, 544
- Kazanskii, A. G., Richards, P. L., & Haller, E. E. 1977, Appl. Phys. Lett., 31, 496
- Love, P. J., Hoffman, A. W., Lum, N. A., et al. 2010, Proc. SPIE, 5902, 590209
- Low, F. J. 1961, JOSA, 51, 1300
- Mainzer, A. K., Eisenhardt, P., Wright, E. L., et al. 2005, Proc. SPIE, 5881, 253
- Maloney, P., et al. 2010, Proc. SPIE, 7741, 77410F
- Mather, J. C. 1982, Appl. Opt., 21, 1125
- Mather, J. C. 1984, Appl. Opt., 23, 3181
- Mazin, B. A. 2005, Microwave kinetic inductance detectors. Dissertation (Ph.D.), California Institute of Technology. <http://resolver.caltech.edu/CaltechETD:etd-10042004-120707>

- McLean, I. 2008, *Electronic Imaging in Astronomy: Detectors and Instrumentation* (2nd ed.; New York: Springer)
- Monfardini, A., Swenson, L. J., Bideaud, A., et al. 2011, *A&A*, 521, A29
- Moore, G. 1965, Cramming more components onto integrated circuits. *Electronics*, 38(April 19), 114–117
- Moseley, S. H., Allen, C. A., Benford, D., Dowell, C. D., Harper, D. A., Phillips, T. G., Silverberg, R. F., & Staguhn, J. 2004, *Nuclear Instruments and Methods in Phys. Res. A*, 520, 417–420
- Petroff, M. D. & Stapelbroek, M. G. 1986, U.S. Patent No. 4568960
- Petroff, M. D., & Stapelbroek, M. G. 1984, Responsivity and noise models of blocked impurity band detectors, in *IRIS Specialty Group on IR Detectors*, Seattle, WA, August 15, 1984
- Pettit, E., & Nicholson, S. B. 1922, *ApJ*, 56, 295
- Planck HFI Core Team et al. 2011, *A&A*, 536, A4
- Poglitsch, A., Waelkens, C., Geis, N., Feuchtgruber, H., et al. 2010, *A&A*, 518, L2
- Ressler, M. E., Cho, H., Lee, R. A. M., et al. 2010, *Proc SPIE*, 7021, 7021O
- Richards, P. 1994, *J. Appl. Phys.*, 76, 1
- Rieke, G. H. 2003, *Detection of Light from the Ultraviolet to the Submillimeter* (Cambridge, UK: Cambridge University Press)
- Rieke, G. H. 2007, *Annu. Rev. Astron. Astrophys.*, 45, 77
- Rieke, G. H. 2009, *Exp. Astron.*, 25, 125
- Rosenthal, D., Beeman, J. W., Geis, N., et al. 2002, *Far-IR, Sub-mm & MM Detector Technology Workshop*, held 1–3 April 2002 in Monterey, CA. Organized and Sponsored by NASA/Ames & USRA/SOFIA. Online at http://www.sofia.usra.edu/det_workshop,id.9
- Rossington, C. S. 1988, Germanium blocked impurity band far infrared detectors. PhD Thesis, Lawrence Berkeley Laboratory, University of California, Berkeley. LBL-25394
- Sclar, N. 1984, *Prog. Quantum Electron.*, 9, 149
- Smith, E. C., Rauscher, B. J., et al. 2009, *Proc. SPIE*, 7419, 741907
- Staguhn, J. G., Benford, D. J., Allen, C. A., et al. 2008, *Proc. SPIE*, 7020, 702004
- Stansberry, J. A., Gordon, K. D., Bhattacharya, B. et al. 2007, *PASP*, 119, 1038
- Szmulowicz, F. & Madarsz, F. L. 1987, *J. Appl. Phys.*, 62, 2533
- Wang, J.-Q., Richards, P. I., Beeman, J. W., et al. 1986, *Applied Optics*, 25, 4127
- Watson, D., & Huffman, J. E. 1988, *Appl. Phys. Lett.*, 52, 1602
- Young, E. T. 1983, *Adv. Space Res.*, 2, 59
- Young, E. T. 2000, *Proceedings of the Space Astrophysics Detectors and Detector Technologies Conference held at the STScI, Baltimore, 26–29 June 2000*
- Zmuidzinas, J. 2012, *Ann. Rev. Condens. Matter Phys.*, 3, 169

15 Astronomical Spectrographs

Rebecca A. Bernstein¹ · Stephen A. Shectman²

¹Astronomy and Astrophysics Department/UC Observatories, UC Santa Cruz, Santa Cruz, CA, USA

²Observatories of the Carnegie Institution, Pasadena, CA, USA

1	<i>Introduction</i>	588
2	<i>Key Concepts in the Design of a Spectrograph</i>	588
2.1	Obtaining the Desired Angular Dispersion	589
2.2	Limitations Due to Pupil Magnification	590
2.3	Resolution, Wavelength Coverage, and Camera Performance	590
3	<i>Optical Components</i>	591
3.1	Collimator	591
3.2	Dispersing Elements	591
3.3	Cameras	594
4	<i>Echelle Spectrographs</i>	596
5	<i>Wide-Field, Multi-object Spectrographs</i>	602
6	<i>Integral Field Spectrographs</i>	608
7	<i>Near-IR Spectrographs</i>	609
8	<i>Spectrographs for Extremely Large Telescopes (ELTs)</i>	611
8.1	Echelles	612
8.2	Wide-Field Optical Spectrographs	613
8.3	IR Spectrographs	614
	<i>References</i>	614

Abstract: While the basic components in optical and near-IR spectrographs have not changed significantly in the last 50 years, spectrograph design and fabrication have become significantly more challenging as telescope apertures grow and the performance goals for the spectrographs become more ambitious. In this review, we discuss the basic optical layout and components of modern astronomical spectrographs and review the designs that have been employed for low-resolution (imaging) spectrographs and high-resolution (echelle) spectrographs on modern telescopes. We begin with a discussion of strategies for optical layouts, collimator and camera designs, and the common dispersing elements in use today. Finally, we discuss the challenges associated with extending these designs to the next generation of instruments needed for the extremely large telescopes being undertaken today.

Keywords: Echelle spectrographs, Fiber spectrographs, Imaging spectrographs, Infrared spectrographs, Integral field units, Spectrographs, Wide-field spectrographs

1 Introduction

There are a variety of ways of thinking about spectrograph performance and design. Reviews exist in the literature that cover a very broad range of relevant topics from the overall function of spectrographs and their use to the general scaling laws of single-object spectrographs for a wide range of telescope diameters (e.g., Bowen 1964). A variety of texts also exist which provide complete references on the basic optical components of spectrographs, including expressions for grating dispersion, angle effects, and optical efficiency (e.g., Schroeder 2000). For the most part, we have tried to avoid repeating those discussions in this review, but concentrate instead on the current strategies being employed in the design of high-performance spectrographs on the current 8 m-class and next generation of even larger telescopes, and the considerations that motivate these designs.

We focus on three general categories of spectrographs currently in use. The first is moderate to high resolution, cross-dispersed spectrographs (echelles and echellettes). These are almost exclusively single-object spectrographs in which cross dispersion allows the spectrograph focal planes to be filled with multiple orders rather than multiple objects. The second category is low to moderate resolution spectrographs. In the last 20 years, observational programs at these resolutions have focused on multiplexing for efficient surveys, so that the workhorse spectrographs on large telescopes at these resolutions are now wide-field, imaging spectrographs. The third category we discuss is near-IR spectrographs, which share the same optical design principles as their optical counterparts, but require special considerations to avoid the effects of high thermal backgrounds. In discussing these instruments, we focus on the general features of the optical designs that have been successfully executed to date or proposed for the next generation of telescopes. Specific instruments are discussed as representative or noteworthy examples within these categories; we do not attempt to present an inclusive record of all successful instruments.

2 Key Concepts in the Design of a Spectrograph

A standard single-object, low-resolution astronomical spectrograph includes five basic components. The first is a slit, or perhaps a fiber, which isolates light from an object in the focal plane of the telescope and rejects the surrounding sky background. The width of the slit defines the size

of a wavelength resolution element in the final spectrum. The slit width is typically matched in size to the resolution of the telescope (or seeing disk) in order to maximize the light collected from the object, the contrast of the object relative to sky background, and the wavelength resolution. The second component is a collimator, which renders the diverging cone of light that emerges from any location in the slit into a parallel (collimated) beam. The third component is the dispersing element (or elements), which alters the angle of propagation of the collimated beam as a function of wavelength. The fourth is a camera, which focuses the angularly dispersed collimated beams into a spatially dispersed image in the spectrograph focal plane. The last is a detector, which measures the intensity of the dispersed image as a function of position in the focal plane, where the position in one direction now corresponds to the spatial location along the slit and the position in the other direction corresponds to wavelength.

For a wide-field spectrograph with a long slit or multiple entrance apertures distributed over a field of view, the collimated beams from different points in the field of view emerge from the collimator at different angles. The narrowest point in the ensemble of collimated beams is the location where the collimator forms an image of the telescope entrance aperture (usually the primary mirror); this image is the collimator exit pupil. While the dispersing elements can be placed anywhere in the collimated beam, it is desirable to locate them as close as possible to the pupil image in order to minimize their size as well as the size of the spectrograph camera.

For seeing-limited spectrographs on large telescopes, the goals of achieving higher spectral resolution, broader wavelength coverage, and more efficient multiplexing all drive designs to be larger – to have larger pupils and larger fields of view, while the limitations on cost, mechanical complexity, and availability of optical materials and coatings all drive designs to be smaller. There are also constraints that involve basic physical and geometric considerations that impose more fundamental limits to achieving high spectral resolution, wavelength coverage, and image quality. The optimization of a spectrograph design typically begins with the consideration of these more fundamental, physical issues, which we discuss below.

2.1 Obtaining the Desired Angular Dispersion

In most cases, the dispersing element will be a diffraction grating. The intrinsic resolving power provided by a reflection grating is given by

$$R = (2/\theta) \tan\delta (d_{\text{coll}}/D_{\text{tel}}) = (2/\theta) \sin\delta (L_{\text{grat}}/D_{\text{tel}}),$$

in which R is the resolving power ($\lambda/\Delta\lambda$), θ is the angular slit width on the sky in radians, δ is the blaze angle of the grating, d_{coll} is the diameter of the collimated beam (or more properly of the collimator exit pupil), D_{tel} is the diameter of the telescope, and L_{grat} is the length of the grating illuminated by the collimated beam. Strictly speaking, this formula is exact only when the grating is used in the Littrow configuration (see below), but it is approximately correct for a much broader range of applications.

To achieve the same wavelength resolution at the same angular slit width on the sky, θ , a spectrograph on a 10 m telescope that employs the same design strategy and grating blaze angle as one on a 4 m telescope will need to be scaled up so that its collimated beam diameter is 2.5 times larger. In practice, this scaling requirement was partly alleviated in going from spectrographs for 4 m-class telescopes to spectrographs for 8 m-class telescopes by incorporating active optics systems in the recent generation of 8 m-class telescopes and building them at the

best available sites; the resulting improvements in seeing-limited performance of 8 m-class telescopes relative to the previous generation have made it fairly routine to use slits as narrow as 0.7 arcsec (as apposed to 1–1.5 arcsec) without incurring unacceptable slit losses.

The need to increase the beam diameter can be further alleviated if the *effective* slit width can be reduced. This might be accomplished by using an optical or a fiber-optic image slicer that reformats a short and wide slit into a longer and narrower one. Note, however, that the total slit area cannot be reduced without violating the conservation of optical intensity. In the near-infrared, further improvement in image quality over significant fields of view may be possible, for example, by using ground-layer adaptive optics (GLAO). But the limitations imposed by the coherence properties of atmospheric turbulence limit the practical application of adaptive optics (AO) correction over any field of view to wavelengths longer than about 8,000 Å.

Finally, note that the scaling of spectrographs with telescope size specifically does not apply when the telescopes are operated at the diffraction limit using adaptive optics. In this case, the image size (θ) decreases with increasing telescope aperture (D_{tel}), and in principle the same spectrograph will produce the same resolution independent of telescope size.

2.2 Limitations Due to Pupil Magnification

In a wide-field spectrograph with a field of view on the sky of ϕ , the range of angles between collimated beams from different field positions at the collimator exit pupil are increased by the ratio of the telescope diameter to the collimated beam diameter ($\phi' = \phi D_{\text{tel}}/d_{\text{coll}}$). This effect is called pupil magnification. For example, a spectrograph with an 80 mm pupil on an 8 m telescope has an angular magnification at the pupil of 100, so that the collimated beams from a 10 arcmin (0.167°) field of view fill a 16.7° cone at the pupil.

For a given field of view on the sky, the optical design problem for the collimator tends to become easier for larger collimated beams with smaller pupil magnification. A smaller range of incident angles onto the grating also tends to result in higher and more uniform efficiency and dispersion. This is one of the motivations for allowing the pupil to become larger, rather than simply using larger blaze-angle gratings to achieve the same resolution.

2.3 Resolution, Wavelength Coverage, and Camera Performance

A fundamental conflict exists in spectrograph design between the desire to increase the spectral resolution and wavelength coverage of a spectrograph, and the practical limitations on the optical performance of spectrograph cameras. The performance of a given camera design might be characterized by its field of view, spatial resolution, the wavelength range over which good images are formed, and efficiency. The complexity of a camera design might be characterized by the number of optical elements or the number of aspheric surfaces it requires. In general, improvements in performance must be paid for with increases in complexity and cost. Practical limits exist in these characteristics beyond which feasible designs are unlikely to exist.

Field of view is a particularly critical characteristic of spectrograph cameras. A given camera has a field of view that can be expressed as the linear diameter of the focal plane over which it produces acceptable images. The angular diameter of the field of view is the diameter of that focal plane divided by the camera focal length. The difficulty of the optical design problem for a spectrograph camera is increased because of the requirement of an external pupil stop

where the grating can be located; the problem becomes worse as more pupil relief – a larger distance between the pupil and the first element of the camera – is required in order to provide enough space for the grating. The highest-performance cameras used in spectrographs for 8 m-class telescopes have angular fields of view in the range 20–25°. Many groups have expended considerable ingenuity in designing these cameras, and the prospects are poor for appreciably increasing the angular field of view that can be obtained.

The angular field of view of the camera must be large enough to accommodate the desired number of spectral resolution elements across one order of the final spectrum. For large telescopes, this angle becomes large quickly because each spectral resolution element spans an angle equal to the width of the slit on the sky multiplied by the pupil magnification. In a wide-field spectrograph, part of the angular field of view of the camera must also be allocated to the angular field of view of the spectrograph on the sky. In order to obtain uniform spectral coverage over the full spectrograph field of view, the field of the spectrograph camera must be large enough to cover a rectangle with a length equal to the length of the spectrum plus the projected field size in the dispersion direction, and a width equal to the projected field size perpendicular to the dispersion.

The optical design problem for spectrograph cameras also becomes more difficult as the focal ratio (f-number) decreases. As seeing-limited spectrographs scale up for larger telescopes, the size of the detector or detector mosaic scales up as well. For practical camera focal ratios and typical detector pixel sizes (both optical and infrared), the pixels are usually small enough to adequately sample the slit width. Camera f-ratios and image quality are then critical to achieving efficient use of the detectors and the best spectrograph performance.

3 Optical Components

3.1 Collimator

In any spectrograph, the collimator design must provide sufficient access to and space around the pupil for the dispersing elements to be located in the narrow portion of the collimated beam. For single objects or small fields of view (<1 arcmin), both refracting and reflecting collimators can provide sufficient image quality and a sharp pupil on any telescope configuration. Refracting collimators have some appealing advantages over reflecting collimators in that they can form an exit pupil that is well controlled and externally located without requiring an off-axis configuration or internal obscurations. They can also be packaged more conveniently and can potentially provide better image quality. The more difficult problem of designing collimators for wide-field spectrographs is discussed in [▶ Sect. 5](#).

3.2 Dispersing Elements

The choice of dispersion strategy is central to the design of any spectrograph. The oldest dispersing element is the prism. Light enters and exits a prism at two flat surfaces tilted with respect to each other by the apex angle. The combined refraction at the two surfaces changes the direction of propagation by an angle called the deviation angle, which is a function of the wavelength-dependent index of refraction, $n(\lambda)$, of the prism material.

For modern spectrographs, the principal virtue of a prism is that it can have very high transmission as long as the required dispersion is modest. For this reason, prisms are often used as efficient dispersing elements in very low-resolution spectrographs, or as cross dispersers in echelle or echellette spectrographs. However, because the index of refraction is not a very strong function of wavelength, the angular dispersion of a prism (the change in deviation angle with wavelength) is limited. Increasing the apex angle can increase the dispersion, but at some point the angle of incidence or the angle of refraction at the tilted surfaces becomes very high, and antireflection coatings become ineffective. If there is sufficient space in the optical train, multiple prisms can be used in series to increase the dispersion without having to resort to impractical angles of incidence. Note also that the index of refraction of optical materials, and hence the wavelength stability of a prism spectrograph, depends on temperature.

The most dispersive glasses tend to have poor transmission in the blue or ultraviolet, so that very large prisms or stacks of prisms can have significant transmission losses through the bulk material. In general, the optical homogeneity of the prism material will not be perfect, and wavefront errors of a collimated beam propagating through the inhomogeneous material can result in significant aberrations in the final image.

A prism is said to be operating at minimum deviation if it has been rotated so that light at a reference wavelength enters and exits at the same angle; at any other orientation the deviation angle will increase. A pair of prisms, one made from a more-dispersive glass and one made from a less-dispersive glass, can be combined to make a *zero-deviation* prism at some reference wavelength. Prisms made from two different materials can be combined to produce a net angular dispersion with significantly different and, potentially, more uniform wavelength dependence than can be achieved with a single material.

The circular cross section of a collimated beam is preserved when propagated through a prism at minimum deviation, but will emerge with an elliptical cross section from a prism where the entrance and exit angles are different. This effect, called anamorphic beam distortion, can be significant and can be used to manipulate the optical properties of a spectrograph in favorable ways. The angular magnification of a beam with anamorphic distortion is different along the major and minor axes of the elliptical beam profile.

Diffraction gratings were developed in order to achieve greater uniformity, greater stability, and higher angular dispersion angular dispersion than is possible with prisms. The basic principle of a grating is that positive interference occurs for certain angles of propagation where there is an integer number of wavelengths between regularly spaced features in the grating. The grating equation gives the relationship between wavelength, λ , and dispersion angle, θ , for a given groove spacing, d : $m\lambda = d \sin\theta$, in its simplest form. The diffraction order is the number of wavelengths, m .

Until recently, astronomical gratings have almost always been surface relief gratings, which are most often made by cutting equally spaced parallel grooves in a soft aluminum coating using a diamond stylus. The aluminum coating is generally deposited on a thick optical flat made from a zero-expansion material that is insensitive to thermal or mechanical deformation. A master grating can be replicated onto another substrate coated with a thin layer of resin. The resin is left transparent in order to produce a transmission grating and is aluminized in order to produce a reflection grating.

In order to achieve high efficiency, a diffraction grating must concentrate most of the diffracted light at a given wavelength into a single order. A grating with this property is said to be *blazed*. For a blazed reflection grating, the grooves are cut at an angle such that the geometric angle of reflection from the surface of each groove is equal to the angle of diffraction in

the desired order. For a blazed transmission grating, the grooves act as tiny prisms for which the deviation angle is equal to the diffraction angle. The angle at which the grating facets are cut relative to the overall plane of the grating is called the blaze angle. For grooves cut at a given angle, the reflection angle of a mirror is much larger than the deviation angle of a prism. For this reason, reflection gratings can be efficiently blazed at higher angular dispersion than is possible for transmission gratings.

To achieve high blaze efficiency, the groove spacing d must not be too small since the blaze effect essentially operates in the geometric optics limit at each grating facet. The profile of the grooves must also be controlled. Gradual wear of the diamond stylus during the ruling process limits the size of gratings that can be produced as a single ruling to about 1,000 cm².

The angle of incidence onto a grating is critical to its performance. Consider a reflection grating in which the collimated beam is incident to the overall surface of the grating at some angle. The undispersed “zeroth order” of diffraction will appear at an equal angle on the other side of the overall surface normal, just like a reflection from a flat mirror. The blaze angle of the facets can be arranged in order to reflect light back toward the incident beam or beyond the reflected (zero order) beam. The first case is referred to as using the interior orders of the diffraction grating, and the second case is referred to as using the exterior orders. When the diffracted light comes back along the same path that the incident light goes in, the grating is said to be operating in the Littrow condition. Just as is the case for a prism, diffraction of the beam at some angle other than Littrow will result in an anamorphic distortion of the cross section of the diffracted beam. If the beam is diffracted outside of the angle between the incident beam and the zero-order beam, it will contract in one direction, while if it is diffracted inside the angle between the incident beam and the zero-order beam it will expand.

A similar condition applies to a transmitting grating, which is said to be operating in Littrow when the overall plane of the grating bisects the angle between the incident and the diffracted beams. The overall optical deviation angle of a transmission grating can be controlled by bonding the transmission grating to or replicating it directly onto a prism. Such a combination is often referred to as a grism.

Volume-phase holographic (VPH) gratings (Barden et al. 2000) are a more recent innovation and are formed by modulating the index of refraction in a thin layer of transmissive material, typically dichromated gelatin. Permanent changes in the index of refraction of this material can be created by exposure to light, using a laser-generated interference pattern with the desired linear fringes. VPH gratings are blazed when they satisfy the Bragg condition for constructive interference of the reflections from the internal planes defined by the index variations. VPH gratings can be made with higher line densities than surface-relief gratings and can exhibit very high blaze efficiencies. The blaze wavelength of a grating can be tuned by changing the angle at which the grating is illuminated. In principle, very large gratings can be fabricated if sufficiently large optics are available to produce the laser interference pattern.

Immersion gratings are surface-relief gratings in which the diffraction occurs inside of a refracting medium instead of in air. The advantage of an immersion grating is that the size of a grating facet that spans a given number of wavelengths is reduced by a factor of the index of refraction, n , and hence the angular dispersion is increased by the same factor. A conventional reflection grating can be used as an immersion grating if the substrate is transparent. In this case, the refracting medium is the replicating resin, which typically has an index of refraction $n \sim 1.5$. Alternatively, the grooves can be etched directly onto crystalline substrates such as silicon using photolithographic techniques. In the case of a silicon immersion grating used in the infrared (where silicon is transparent), the index of refraction is very high ($n \sim 3.45$), and the size of a

grating and of a grating spectrograph with a given wavelength resolution can be reduced by a corresponding factor, which is large enough to be highly advantageous, especially in a cryogenic instrument (Jaffe et al. 2006, 2008).

An echelle grating is a coarse reflection grating that is operated in very high order, typically near 100. When operated near Littrow, the overall surface of an echelle grating is tilted at an extreme angle to the incident and diffracted beams, typically as high as $60\text{--}80^\circ$. The tangent of this angle is called the R-factor of the grating (the tangent of the blaze angle, $\tan \delta$). An R2 grating ($\tan \delta = 2$), for example, must be approximately twice as long as it is wide in order to intersect all of the light from a circular beam. The angular dispersion of a grating operated in Littrow is proportional to the R-factor, so using such very oblique gratings is one way to achieve high spectral resolution.

It is a general property of all of the dispersing elements discussed in this section that the optical path length for a circular beam propagating through the system changes by a large amount from one side of the beam to the other. This is the condition that requires light incident on the prism or grating to be accurately collimated; if the light is not completely parallel, then one side of the beam will travel further and will be more out of focus than the other side. In this way, the dispersing element will efficiently convert any focus error of the collimator into coma in the final image, which cannot be compensated by a simple focus offset of the spectrograph camera.

3.3 Cameras

Because most of their optical power comes from reflective surfaces, catadioptric cameras (mixed lens and mirror systems) such as Schmidt cameras have the advantage that they can be readily achromatized, even well into the ultraviolet and at very fast focal ratios. However, it is hard to design catadioptric cameras with fields of view larger than about 15° . The principal limitation on the achievable field of view arises from the requirement that the detector is flat. This has led some groups to investigate the feasibility of fabricating detector arrays with appreciable net curvature (e.g., Dumas et al. 2010). For single-object, cross-dispersed spectra, a 15° field of view is often sufficient, and generalized Schmidt cameras have been used successfully in echelle and echellette spectrographs such as HiRES on Keck or MagE on Magellan. Although cameras of this type often have only one reflection and a few air/glass surfaces, they suffer from a central obscuration that generally blocks 20% or more of the incident light, depending on the strategy employed for packaging the detector at an internal focal plane or extracting the image to an external location. Off-axis designs are sometimes used to avoid the obscuration, but only with a substantial penalty in optical performance.

Dioptric (lens only) systems have the advantage that they avoid the central obscuration that is characteristic of Schmidt cameras. Refracting cameras can be made compact by placing elements with substantial optical power as close as possible to the dispersing elements. Good image quality can be obtained over wide ranges in wavelength ($0.4\text{--}1.0\ \mu\text{m}$) with careful selection of glass types to obtain apochromatic correction over the desired wavelength range. Apochromatic correction typically involves the use of elements made from crystalline calcium fluoride (CaF_2) or CaF_2 -like glasses (e.g., S-FPL51, made by Ohara Corp.), which have very low dispersion and thus can provide a lot of power with minimal chromatic aberration relative to typical optical glasses.

To achieve high optical efficiency, antireflection (AR) coatings are typically applied to all air/glass surfaces. Multilayer dielectric coatings can be used to keep reflection losses as low as

0.5–1.0% per surface over bandpasses of more than an octave (e.g., 400–1,100 nm) and over an appreciable range of incident angles. The high-order interference of multilayer coatings typically goes bad rapidly outside of the design wavelength range. For even wider bandpasses, better performance is obtained overall by using single-layer AR coatings, although these will have losses of up to 2% per surface at the ends of the bandpass. Some improvement has been achieved with new hybrid coatings such as SolGel + MgF₂ that may be more widely available in the future.

Alternatively, lenses can be bonded in multiplets to reduce the reflection losses. Hard optical epoxies and cements are usually avoided when bonding CaF₂ or CaF₂-like materials (which have much higher coefficients of thermal expansion) to other types of glass. Compliant optical cements or RTVs have been used successfully for this purpose. Such RTVs have shear compliance that is adequate to take up the differential thermal expansion coefficients of the bonded materials, as long as the bond layer is adequately thick and the surfaces are not too strongly curved. An additional advantage of bonding lenses with RTV is that delicate or humidity-sensitive optical materials (e.g., CaF₂, BaF₂, or NaCl) can be permanently isolated in multiplets where the external materials are more conventional glass. Oil coupling has also been used in several large camera systems with 250–350 mm diameter lenses. In some cases, the oil has even been used to create powered, liquid elements to provide thermal focus compensation (see discussions in Dressler et al. 2011). Because of the extreme range in temperature between the laboratory and in-service environments, bonded or coupled elements are seldom used in lens assemblies for cryogenic instruments.

The fundamental design limitation of dioptric cameras is the wavelength dependence of image quality caused by the change in index of refraction with wavelength. An important component of the optical design problem is therefore the careful choice and deployment of different optical materials in order to compensate for chromatic effects. For this reason, the availability of materials with a wide range of optical properties, and in large enough sizes, is critical to the design of high-performance spectrograph cameras for large telescopes. Certain optical materials can be produced in very large sizes (>1 m diameter × 0.5 m thick) and weights (hundreds of kg, Jedamzik and Hartmann 2006). However, the most readily available materials in these sizes tend to cover a small area in the plot of index of refraction vs. wavelength that characterizes different glass types. A wider range of glass types is available in blank sizes up to roughly 350 mm diameter × 100 mm thick. Many very useful glass types are restricted to even smaller sizes. Some glasses can be slumped after annealing to produce meniscus lens shapes without having to produce an initial cylindrical blank that encompasses the full dimensions of the final lens.


High-quality optical glass is available from only a few vendors and the availability of particular types of glass in particular sizes changes with time. Typically, it is necessary for the designer to check the availability of different types of glass in the sizes required, making the design process highly iterative. In some cases, the limitations on the availability of optical glass blanks of a given size are physical, such as the allowable cooling time for the glass melt before the mixture of components separates at a microscopic scale (called devitrification) or the difficulty of adequately annealing large blanks without causing unacceptable stress and breakage. It is not evident whether there are significant opportunities to expand the limits of the current production methods for these materials.

Availability is even more limited for materials with good transmission at wavelengths blueward of 375 nm. The only materials that will not attenuate significantly down to the atmospheric cutoff (300–320 nm, depending on altitude) are calcium fluoride and fused silica. Because of its high transmission and low dispersion, calcium fluoride is a critical material for cameras

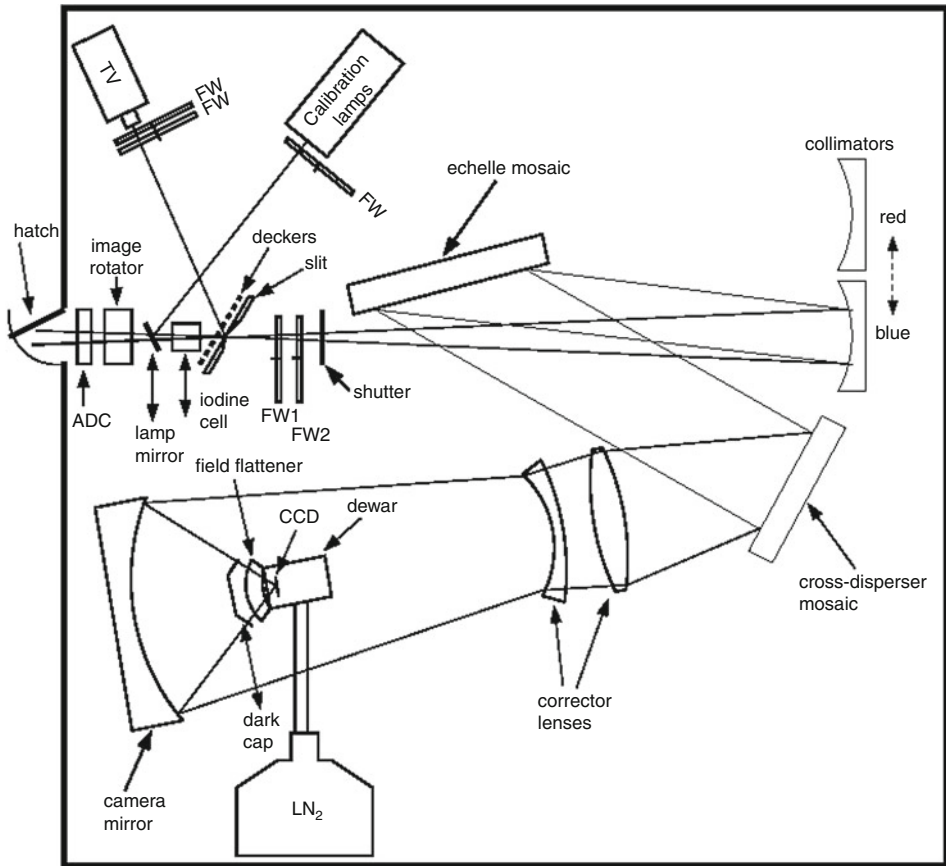
operating at any wavelength. Its availability at large diameters (>350 mm) and thicknesses (>50 mm) is therefore critical to the continued use of dioptric cameras on large telescopes. While the availability of large boules has come and gone over the years, driven most recently by the photolithography industry for semiconductor fabrication, there is no fundamental physical limitation to growing larger crystals, and several companies are developing production facilities aimed at producing 500 mm diameter \times 100 mm thick blanks.

4 Echelle Spectrographs

Modern high-dispersion spectrographs achieve high resolution by using echelle gratings, which have modest line density (30–80 lines/mm) but are blazed at very large angles ($\tan \delta \geq 2$) and used in high orders of diffraction, m . The wavelength spacing between orders, or the free spectral range, is equal to λ/m . At this full width, the blaze intensity function for a single order of a high-quality echelle ruling falls to about 50% of its peak value, so that most of the light at a given wavelength is concentrated in a single order, or is split between at most two adjacent orders. Typically, a second dispersing element is used to weakly disperse the light in the direction perpendicular to the echelle orders, so that the orders do not overlap and complete coverage over the desired bandpass can be obtained in a single exposure on a rectangular detector. Cross-dispersion is not necessary if an order-blocking filter is used to transmit a single order, as is sometimes the case when fibers are used to obtain a single order from multiple objects rather than multiple orders from a single object. Illustrative examples of such multi-object instruments include Hectochelle (Szentgyorgyi et al. 2011a) on the MMT, FLAMES + Giraffe and FLAMES + UVES on the VLT (Pasquini et al. 2002), and MMFS on Magellan (Walker et al. 2006).

Echelle spectrographs for 8 m-class telescopes employ several different design strategies, with different advantages and disadvantages. The standard layout for a reflecting-grating spectrograph uses the grating tilt in the plane of the dispersion to send the diffracted light toward the camera and away from the incoming beam from the collimator. When this arrangement is used with the very high blaze angle of an echelle grating, the anamorphic beam expansion can be as high as a factor of 2 or more. This is true even when the angle between the collimator and the camera is kept as narrow as possible in order to minimize the anamorphic effect. The desirable consequence is that the angular slit width is reduced by the anamorphic factor so that the resolution is correspondingly higher, but the undesirable consequence is that the entrance pupil of the camera and thus the camera itself must be larger, and the grating is even longer than would be calculated from the factor $\tan \delta$. The HiRES spectrograph on Keck is the most ambitious and successful execution of this basic configuration (Vogt et al. 1994; see  Fig. 15-1). It uses a second reflecting grating for cross dispersion which requires even more pupil relief and further increases the size of the camera. The collimator for HiRES is a mirror, and the camera is a “super Schmidt” design with significant obscuration by the detector and its associated packaging. The exit pupil of the HiRES collimator is 300 mm in diameter, and the echelle grating is a mosaic of three replica gratings aligned end to end in a mechanical mounting. Even the cross disperser in this very large instrument is a mosaic of two replica gratings.

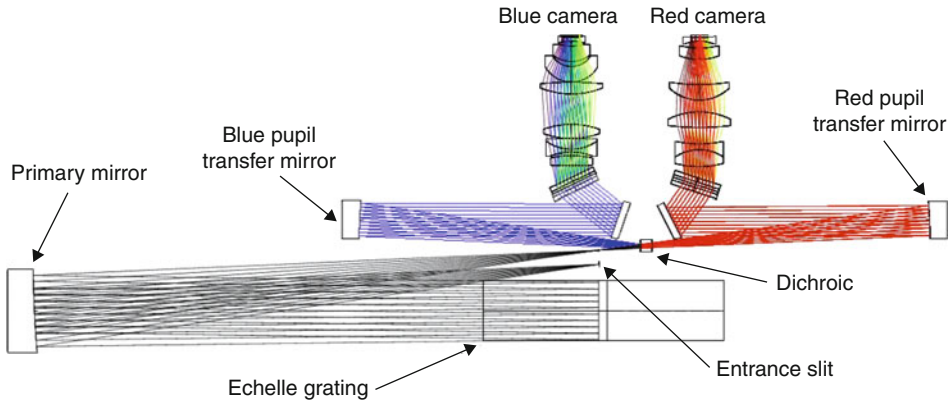
Another successful design strategy is called a “white-pupil” configuration, in which the echelle grating is used at a near-Littrow angle (see Baranne et al. 1974 and references therein). In such a design, the light from a slit is collimated using a mirror, and the diffracted light from



■ Fig. 15-1

A simplified, functional schematic of the optical layout of HiRES on Keck is shown to illustrate the standard configuration for an echelle spectrograph (see Vogt et al. 1994). This graphic is a version of the xhires user interface used by observers at Keck to configure the instrument (see <http://www2.keck.hawaii.edu/inst/hires>). Note the anamorphic magnification that occurs at the echelle grating (labeled “echelle mosaic”). The echelle grating is tilted about an axis that is perpendicular to the page in this sketch. Although it is not apparent from the drawing, the cross-dispersing grating (labeled “cross-disperser mosaic”) is tilted about an axis that is parallel to the page, so that anamorphic magnification occurs there as well. This can be used to obtain higher dispersion, as discussed in the text

the grating is directed back to the same mirror, which then re-forms an image of the slit, now linearly dispersed by the echelle grating. As long as the slit is fairly short, the slit and its dispersed image are both narrow, and a small offset of the grating angle is all that is required to keep the two separate. The light then diverges from the dispersed image and is re-collimated by a second mirror. The second collimator forms an image of the echelle grating, which is an undispersed (white) pupil. A cross-dispersing grating or prism is placed at this second pupil



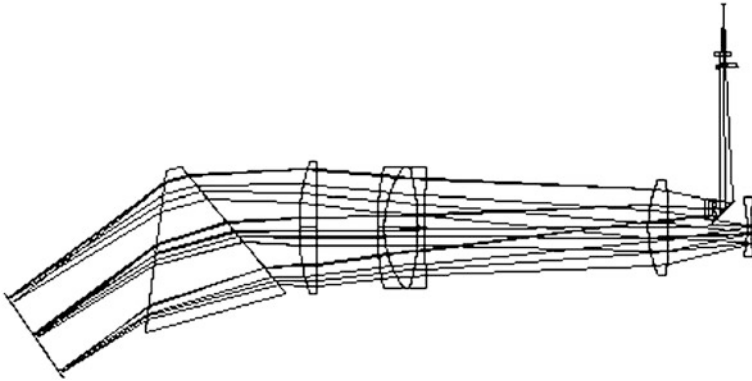
■ Fig. 15-2

The optical design for the SALT HRS spectrograph is shown to illustrate a white-pupil echelle configuration. The key elements in this configuration are particularly well illustrated in this figure from Barnes et al. (2008). Light comes from the slit near the center of the drawing and is collimated by the first collimating mirror (labeled “primary mirror”). After it is dispersed by the echelle grating, the light is refocused to form a dispersed image of the slit at a location shown just above the slit in the figure. The beam is then split by a dichroic into a red and blue channel. The mirrors labeled “pupil transfer mirror” recollimate the beam in each arm, this time at a smaller diameter. Flat mirrors fold the beams before they are cross-dispersed by VPH gratings

image, and the light is subsequently directed to the camera and detector. An example of this layout is shown in ● Fig. 15-2.

The feature that cross dispersion can occur at the second pupil has several very nice implications for the optical design of the spectrograph components. First, it minimizes the pupil relief required at the camera because only the cross disperser must be located in the beam between that second pupil and the camera. This allows the camera to be smaller. Also, the second collimator does not have to be the same focal length or produce the same collimated beam (pupil) diameter as the first; lower dispersion is required from the cross disperser and can be achieved with a smaller grating and a smaller diameter pupil. This again allows the camera to be smaller, although its angular field of view must be larger because the angular free spectral range of the echelle orders increases by the same factor that the diameter of the second pupil decreases. This is another example of pupil magnification, discussed in ● Sect. 2.

The white-pupil arrangement has been used in a number of spectrographs, most notably the UVES spectrograph at the ESO VLT (Dekker et al. 2000). The collimator mirrors must be used somewhat off-axis in order to separate the slit from the re-formed, dispersed slit image. Also, the angular field of view of the collimator must be significant to accommodate the dispersed beam returning from the echelle grating. For this reason, the white-pupil arrangement is generally implemented using fairly slow first and second collimators, with focal ratios of $f/10$ or more. Often, one or more small-fold mirrors are required near the slit or the dispersed slit image in order to create a clear path for light to enter the spectrograph and to place the second collimator and camera at mechanically favorable positions. There can also be significant field curvature effects at the dispersed slit image, which must be explicitly compensated in the design of the camera.



■ Fig. 15-3

The optical design of the red arm of the MIKE spectrograph on Magellan is shown to illustrate an all-refracting, double-pass echelle configuration (The figure is reprinted from Bernstein et al. (2003)). The echelle grating is located at the far left and is drawn here as a flat mirror. Light is shown emerging from the slit at the top of the figure. In order after the slit, the optical elements include a dichroic, a field lens, a fold mirror, a small triplet, the camera elements (singlet-triplet-singlet), a prism, and finally the grating. From the grating, the return beam is slightly offset from the incoming beam to avoid the small triplet on the way to the field flattener and detector at the far right of the figure. To accomplish this offset, the grating is tilted very slightly to move the beam cross dispersion direction, rather than in the direction of the echelle orders. This allows the grating to be used in Littrow. The field lens and small triplet comprise the “injection optics.” See text for further discussion

A third configuration for an echelle spectrograph uses an all-refracting camera in double pass and, like a white-pupil layout, operates very close to Littrow (see ► Fig. 15-3). In this configuration, light from the slit is collimated going through the camera on its way to the grating on the first pass, and the image of the spectrum is formed going back through the same camera after being dispersed by the echelle grating. A small tilt of the grating separates the images of the slit and the detector in the camera focal plane. Cross dispersion is accomplished with a prism (or prisms) placed between the camera and the grating, which are also used in double pass. The focal ratio of the camera is selected to provide the desired image scale at the detector, and a set of injection optics must then be used between the telescope focal plane and the camera to adapt the focal ratio of the telescope to the focal ratio of the camera. This design has the advantage of being compact and cost-effective. Although the available cross dispersion using prisms is lower than with diffraction gratings, the higher efficiency of the prisms is also a significant advantage. The double-pass arrangement has been used in a number of spectrographs, most notably in MIKE at Magellan (Bernstein et al. 2003).

The point of building bigger telescopes is to make fainter targets accessible. On 8 m-class telescopes, high-resolution spectrographs designed for high throughput have reached brightness limits near $V \sim 20$ AB mag. Good sky subtraction and low scattered light have also been critical to reaching such faint limits. Respective, these require high quality optical elements and high image quality, and sufficient slit length for accurate measurement of the sky signal.

■ Table 15-1

A representative sample of cross-dispersed optical and near-IR spectrographs (echelles and echellettes) on 8 m-class telescopes. Column 3 indicates the wavelength range for each available channel. Channels can be exposed simultaneously unless otherwise noted. Column 4 gives the *approximate* transmission efficiency of each channel. These values include the telescope unless otherwise noted. Column 5 lists the available slit widths or the range if continuously selectable. The slit width in square brackets provides the resolution listed in Column 9; resolution scales linearly with slit width for all instruments. Column 6 indicates the minimum order separation and the kind of cross dispersion used (prism or grating). Column 7 gives the collimated beam diameter at the echelle grating. Column 8 gives the tangent of the blaze angle (δ) of the echelle grating

	Telescope (diam)	λ -range (μm)	Efficiency	Slit widths (arcsec)	X-disp (arcsec)	Pupil (mm)	$\tan\delta$	R ($\lambda/\delta\lambda$)
HIRES ^a	Keck (10 m)	0.36–0.65 (blue) ^a	0.06–0.15	0.40, 0.57, [0.86], 1.15	6 (g)	300	2.8	49,000
		0.38 – 1.10 (red) ^a	0.02–0.20	0.40, 0.57, [0.86], 1.15	6 (g)	300	2.8	50,000
ESI ^b	Keck (10 m)	0.39 – 1.1	0.12–0.27	0.3, 0.5, 0.75, [1.0], 1.25, 6	20 (p)	150	0.63	4,050
UVES ^c	VLT (8 m)	0.3–0.5 (blue)	0.06–0.12	0.15–[1.0]–2.0	10 (g)	200	4	41,400
		0.42–1.1 (red)	0.03–0.16	0.15–[1.0]–2.0	12 (g)	200	4	41,400
X-Shooter ^d	VLT (8 m)	0.3–0.55 (blue)	< 0.22	0.5, 0.8, [1.0], 1.3, 1.6, 5.0	12 (p)	100	0.9	5,100
		0.55–1.0 (red)	< 0.22	0.4, 0.7, [0.9], 1.2, 1.5, 5.0	12 (p)	100	1.4	8,800
		1.0–2.3 (near-IR)	< 0.2	0.4, 0.6, [0.9], 1, 1.2, 5.0	12 (p)	85	1	5,100

bHROS ^e	Gemini (8 m)	0.4–1.0	0.04–0.14 ^f	0.7, 0.9 ^e	7.2 (p)	175	2	150,000 ^e
Phoenix ^g	Gemini (8 m)	1.0–5.0	0.04	0.17, [0.25], 0.34	14 ^g	200	2	65,000
HDS ^h	Subaru (8 m)	0.3–0.5 (blue)	0.03–0.08	0.2–[1.0]–4.0	4.4 (g)	270	3	32,000
		0.42–0.92 (red)	0.02–0.13	0.2–[1.0]–4.0	4.8 (g)	270	3	32,000
MIKE ⁱ	Magellan (6.5 m)	0.32–0.50 (blue)	0.04–0.35 ^f	0.35, 0.5, 0.7, [1.0], 1.5	5 (p)	150	2.4	28,000
		0.49–1.0 (red)	0.04–0.20 ^f	0.35, 0.5, 0.7, [1.0], 1.5	5 (p)	150	2	22,000
FIRE ^j	Magellan (6.5 m)	0.82–2.52	0.02–0.27 ^f	0.45, [0.60], 0.75, 1.00	7 (p)	50	1	6,000
MaGE ^k	Magellan (6.5 m)	0.30–1.02	0.02–0.22 ^f	0.5, 0.7, 0.85, [1.0], 1.2, 1.5, 2, 5	10 (p)	100	0.63	4,100

^aSee Vogt et al. (1994) and updated information at <http://www2.keck.hawaii.edu/inst/hires>. Red and blue configurations use different cross-dispersing gratings and collimators; one configuration is available at a time. A total of 300–450 nm can be obtained at one time in either configuration. Order-blocking filters are available for use with a long slit (<28 arcsec)

^bSee Sheinis et al. (2002). A low-dispersion (prism), long-slit mode is available. Direct imaging is available over a 2×8 arcmin² field of view

^cSee Dekker et al. (2000). Blocking filters can be used with a long slit (<30 arcsec) to obtain a single order

^dSee Vernet et al. (2011)

^eRetired. See Aderin (2004). The efficiency includes fiber and telescope losses. The output of the fibers passes through an image slicer to feed the spectrograph with the equivalent of a 0.14 arcsec wide slit to obtain the resolution in Column 9

^fValue indicates the efficiency of the spectrograph (slit to detector) only

^gRetired. See Hinkle et al. (1998, 2000, 2002). Orders 11–52 are available with blocking filters to provide one order per exposure

^hSee Terada et al. (2008). Red and blue configurations use different cross-dispersing gratings and collimators; one configuration is available at a time

ⁱSee Bernstein et al. (2003)

^jSee Simcoe et al. (2010) and Simcoe et al. (2008)

^kSee Marshall et al. (2008)

In a cross-dispersed instrument, the slit length competes with the number of orders that can fit across the detector or within the camera field of view and therefore creates a trade-off of slit length against wavelength coverage or resolution. Longer slits also require stronger cross dispersion, which may result in lower efficiency when gratings are required rather than prisms. The point is that slit length can come at a cost, and it is often better to provide only as much slit length as is necessary for a particular observation.

Because they are typically single-object instruments, echelle and echellette spectrographs can easily incorporate a dichroic to split the beam, either immediately before or after the slit (see [Table 15-1](#)). Both the UVES (white-pupil) and the MIKE (double-pass) spectrographs incorporate dichroic beam splitters to send the light to separate blue and red wavelength-optimized spectrographs. The X-shooter spectrograph on the VLT uses a pair of dichroics to send the light to three separate spectrographs for the blue, red, and infrared (Vernet et al. 2011; Spanò et al. 2006).

In the past 15 years, a class of echelle spectrographs has emerged which is highly optimized for precision radial-velocity (PRV) measurements in order to detect and characterize extrasolar planets. These PRV spectrographs require high resolution along with highly accurate wavelength calibration and stability. Observing through an iodine absorption cell has been shown to be capable of producing useful PRV results (Marcy and Butler 1992; Butler et al. 1996) using existing spectrographs (e.g., HiRES at Keck) or purpose-built spectrographs (e.g., PFS at Magellan, Crane et al. 2010). Use of an iodine cell results in some loss of efficiency and is also restricted to the iodine absorption band (500–620 nm).

The HARPS spectrograph at the ESO 3.6 m telescope has shown that it is possible to build a spectrograph with high enough mechanical and optical stability to obtain excellent radial-velocity precision without using an iodine cell (Mayor et al. 2003; Udry et al. 2006). The important technical considerations for obtaining such high stability include temperature control, vacuum isolation of the entire optical train or at least the dispersing elements, and a stable, uniformly illuminated slit and pupil. The most effective way to obtain such highly stable and uniform illumination is using a fiber feed. The fiber feed acts as an image scrambler, either relying on the properties of the fibers themselves or incorporating explicit scrambling optics. Recently, octagonal and hexagonal fiber cores have emerged as an extremely convenient way to obtain even more uniform illumination than can be obtained with cylindrical fiber cores, without requiring any special optics (e.g., Perruchot et al. 2011). A fiber feed can be readily adapted to a compact white-pupil or double-pass design, making it easier to achieve the required thermal stability and to enclose the spectrograph in a vacuum chamber.

5 Wide-Field, Multi-object Spectrographs

Spectroscopic surveys place a premium on the ability to observe many objects at the same time. The field of view of a survey spectrograph should be large in the telescope focal plane, so that a large number of targets will be available at one time, and also in the spectrograph focal plane, so that many objects can be observed with high resolution and wavelength coverage in a single exposure. Thus, the important, and often conflicting, figures of merit for a wide-field spectrograph are field of view, wavelength coverage, resolution, and (of course) efficiency.


There are two principal approaches to wide-field spectroscopy. The first is to use an imaging spectrograph in which the full field of view of a wide-field collimator is reimaged at the focal

plane of the camera after being dispersed by the grating. A custom-made slit mask, with many small slits chosen so that the resulting spectra do not overlap on the detector, is used to select the desired objects and to reject sky. The second approach is to use optical fibers placed at the positions of interesting objects in the telescope focal plane to carry light from each object to a pseudo-slit (made up of a long row of resolved fibers) at the input of a long-slit spectrograph.

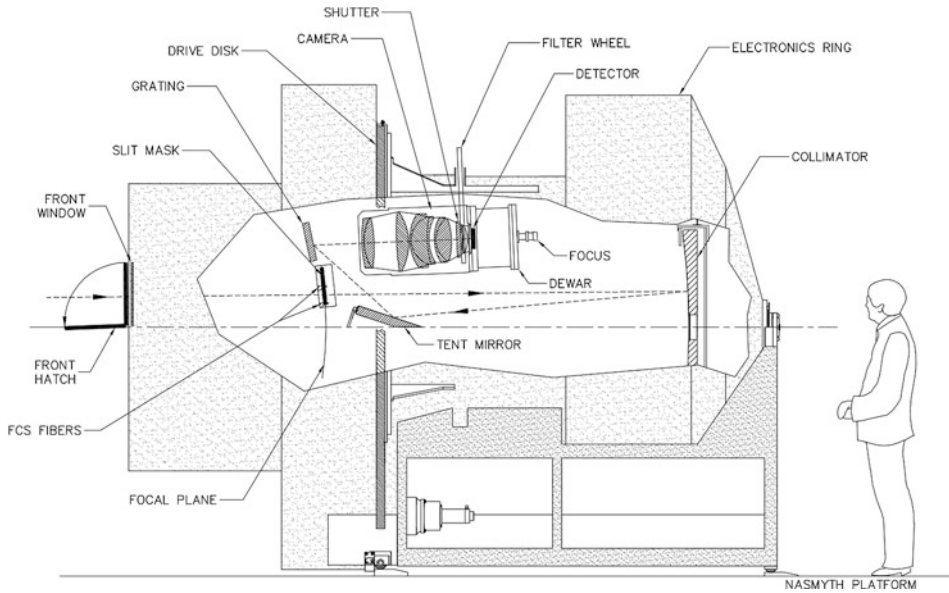
For an imaging spectrograph, the field of view of the camera must be allocated partly to the reimaged field of view on the sky and partly to the spectroscopic dispersion. To the extent possible, it is desirable for the wavelength coverage to be uniform for objects that are distributed across the slit mask, which requires that the angular field of the camera be significantly larger than the angular field of the collimator. In some cases, it is necessary to restrict the size of the slit mask in the dispersion direction, to restrict the wavelength coverage, or to accept that the wavelength coverage will vary to some extent across the field. The size of the slit mask in the direction perpendicular to the dispersion typically determines the number of objects that can be observed at one time.

If the camera is designed to avoid lateral color, then the disperser can be removed from the beam and the spectrograph can be used as an imager. This is often convenient for the purpose of aligning the slit mask on the sky and sometimes for imaging in general, especially with narrow-band interference filters or Fabry-Perot etalons (e.g., TAURUS, Atherton et al. 1982, or more recently MMTF on IMACS, Veilleux et al. 2010).

The optical design problem for the wide-field collimator would seem to be relatively easy because the focal ratio of the telescope, and therefore of the collimator, is relatively slow, typically in the range $f/8$ – $f/15$. In practice, the problem is difficult because of the need to simultaneously control the field curvature and the location of the exit pupil. Across a sufficiently wide field of view, the effect of curvature of the telescope focal plane, and of the collimator itself, is generally significant. It is not possible to place an arbitrary field flattener near the focal plane of the telescope because the field flattener also acts as a field lens and must produce a pupil image at an appropriate place to put the grating. The problem of designing very wide-field collimators boils down to choosing an appropriate configuration so that any required field lens also produces a pupil image at the desired location while also matching the field curvature of the collimator to the field curvature of the telescope.

For most Cassegrain telescopes, the field of view as seen from the instrument is convex. The field of view of a single mirror used as a collimator is also convex (it is easiest to think of the mirror as a correctorless Schmidt with the usual symmetry about the center of curvature). In designing the LRIS and later the DEIMOS spectrographs at Keck (Epps 1990; Faber et al. 2003), Epps realized that the field curvature of the telescope and of a reflecting collimator approximately matches if the collimated beam diameter is about 150 mm (see  Fig. 15-4). In this case, no field flattener is required and the exit pupil appears at the usual position, one focal length away from the collimator mirror. If the optical axis of the collimator mirror is aligned with the optical axis of the telescope, then the collimator field of view must be displaced to an off-axis position in the telescope focal plane, so that the pupil is formed adjacent to the targeted field rather than on top of it.

For a Gregorian telescope, the field of view as seen from the instrument is concave. In general, refracting collimators exhibit symmetry about their exit pupils and have fields that are concave as well, making them particularly well matched to Gregorian telescopes. For refracting collimators, the location of the exit pupil is convenient, and there are more degrees of freedom available to compensate for aberrations over a wide field of view. In designing the Magellan telescopes, Shectman (1994) chose a particular value for the focal ratio produced by the secondary



■ Fig. 15-4

A sketch of the optical layout for the DEIMOS spectrograph on Keck is shown to illustrate a wide field optical spectrograph layout that uses a mirror collimator (The figure is reprinted from Faber et al. (2003)). The slit mask (labeled) indicates the location of the DEIMOS field of view in the curved focal plane of the telescope. Note that the field is offset from the telescope's optical axis, which is indicated by the *long-short dashed line* across the *middle* of the figure. The collimator reflects the beam back toward the focal plane, so that the pupil would form near the telescope focal plane. The tent mirror (labeled) folds the beam to allow access to the pupil, which is located at the grating. As the Keck telescope is a Ritchey-Chrétien design, the focal plane curvature is particularly well suited to a mirror collimator

mirror in order to obtain an exact match to the field curvature of a particular collimator. As a result, the IMACS spectrograph on Magellan has a very wide field of view with high image quality (Dressler et al. 2011; see ► Fig. 15-5) in comparison to other imaging spectrographs.

There are a number of ways to ameliorate the field curvature problem when such an exact match is not available. One way is to explicitly restrict the field of view, since the field curvature mismatch grows with the square of the field diameter. Another way is to redistribute the optical power between the elements of a compound collimator assembly in order to alter the value of the field curvature or the position of the exit pupil. These strategies account for the many imaging spectrographs on Cassegrain telescopes with refracting collimators, for example, GMOS on Gemini (Allington-Smith et al. 2002), MMIRS and Binospec on the MMT (McLeod et al. 2004; Fabricant et al. 1998, 2003), and MOSFIRE on Keck (McLean et al. 2010). Imaging spectrographs with reflecting collimators on Gregorian telescopes are also possible, for example, MODS on the LBT (Pogge et al. 2010), although this configuration does not provide the field of view that is potentially available on a Gregorian telescope.

A final strategy to increase the field of view is to deploy multiple collimators (together with their associated gratings and cameras) across the field of a telescope in what is called a fly's-eye

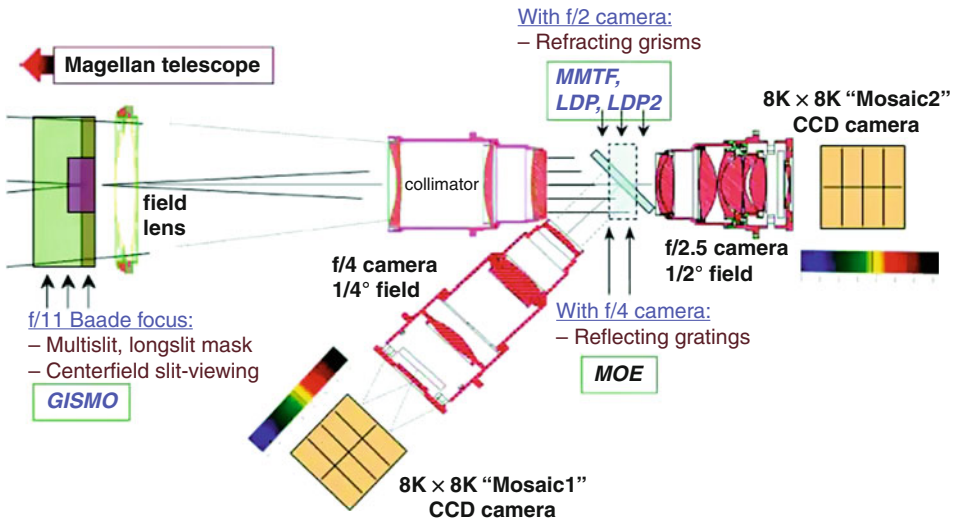


Fig. 15-5

A sketch of the optical layout for the IMACS spectrograph on Magellan is shown to illustrate a wide field optical spectrograph layout that uses a refracting collimator (The figure is reprinted from Dressler et al. (2011)). The IMACS field of view is on-axis in the telescope focal plane. As discussed in the text, the Gregorian configuration of the Magellan telescopes is particularly well suited to this collimator design and allows a very wide field of view (see Table 15-2)

arrangement. The VIMOS spectrograph on VLT contains four such assemblies (Le Fèvre et al. 2003). The designs for DEIMOS and for Binospec include two arms, although the second arm of DEIMOS has not been implemented. In order to deploy such multiple spectrographs in a fully independent manner across the field of view, the corrected telescope focal plane should ideally have the property that the chief ray is always perpendicular to the focal surface. This is not explicitly the case for any of the multiple spectrographs that have been built to date, but Shtetman has proposed a field corrector with this property for use on the Giant Magellan Telescope.¹

Table 15-2 shows the field size, resolution, and wavelength coverage that have been achieved by most of the wide-field spectrographs built for 8 m-class telescopes. The median seeing-limited image quality delivered by all of the 8 m-class telescopes tends to be in the 0.6–0.8 arcsec range, and the spectrographs are all designed to work at least some of the time on faint galaxies which are at most partially resolved. Because the choice of efficient and practical diffraction gratings is also highly constrained, all of the imaging spectrographs tend to deliver resolution that scales in the expected way with the ratio of telescope aperture to collimated beam diameter. Typically, the resolution obtained is $1,000 < R < 5,000$, the wavelength coverage is 370–1,000 nm, and the field area is 25–100 arcmin². Spectrographs with separate red- and blue-optimized channels have better UV response, which can be extended as far as the atmospheric cutoff.

¹See Chaps. 6 and 13.6 of the GMT Conceptual Design Report (<http://www.gmto.org/science-conceptu.html>).

Table 15-2

A representative sample of optical, multi-object, imaging spectrographs on 8 m-class telescopes. Column 3 indicates the wavelength range for all available channels. Channels can be exposed simultaneously unless otherwise noted. Columns 4 and 9 give the *approximate* peak transmission efficiency of each channel for spectroscopy and imaging, respectively. These values include the telescope efficiency unless otherwise noted. Column 5 lists the field of view for imaging; full wavelength coverage is typically available only for slits located within a narrower field of view. Column 6 lists the total field area consistent with Column 5. Column 6 gives the maximum combined slit length possible when multiple slits are used. Column 8 gives the diameter of the collimated beam at the grating. Column 10 gives the resolution obtained using a 1 arcsec slit

	Tel. (diam)	λ -range (nm)	Spec. eff.	FOV (arcmin)	FOV (arcmin ²)	Slit length (arcsec)	Pupil diam (mm)	Imaging eff.	R ($\lambda/\delta\lambda$)
LRIS ^a	Keck (10 m)	310–650 (blue)	0.5	5.5 × 7.5	40	450	150	0.6	300–5,000
		380–1,100 (red)	0.4	5.5 × 7.5	40	450	150	0.6	300–5,000
DEIMOS ^b	Keck (10 m)	410–1,100	0.35	16 × 5.0	80 ^c	980	150	0.5	1,200–8,000
OSIRIS ^d	GTC (10 m)	365–1,000	0.35	8.0 × 5.2	41.6	480	80	0.4	250–2,500
VIMOS ^e	VLT (8 m)	370–1,000	0.4	7.0 × 8.0 [×4] ^e	224 [×4] ^e	480 [×4] ^e	140	0.5	180–2,500
FORS ^f	VLT (8 m)	330–1,100	0.5	6.8 × 6.8	46	408	90	0.5	260–2,600

GMOS ^g	Gemini (8 m)	360–940	0.4	5.5 × 5.5	30	330	100	0.4	670–4,400
FOCAS ^h	Subaru (8 m)	365–900	0.35	6.0 diam.	28	360	90	0.4	500–3,000
MODS ⁱ	LBT (8 m)	320–650 (blue)	0.4	6.0 × 6.0	36	360	230	0.4	100–1,500
		550–1,100 (red)	0.4	6.0 × 6.0	36	360	230	0.4	100–1,700
IMACS ^j	Magellan (6.5 m)	390–1,000 (F/2)	0.4	27 × 27	700	1,620	150	0.5	60–650
		360–1,000 (F/4)	0.4	15 × 15	225	900	150	0.5	500–4,600 [21,000, 0.6 ^l]k

^gSee Oke et al. (1995) and McCarthy et al. (1998)

^hSee Faber et al. (2003)

ⁱThe field of view is vignetted so that this area is approximate

^jSee Cepa et al. (2000). Updated information is available at <http://www.gtc.iac.es/en/pages/instrumentation/osiris.php>

^kSee Le Fèvre et al. (2003). Updated information is available at <http://www.eso.org/sci/facilities/paranal/instruments/vimos>. See also Hammersley et al. (2010) VIMOS has four simultaneously exposed fields of view. The dimensions and area given are for a single field

^lSee Appenzeller et al. (1998)

^mSee Allington-Smith et al. (2002) and Hook et al. (2004)

ⁿSee Kashikawa et al. (2002) and <http://www.naoj.org/Observing/Instruments/FOCAS/index.html>

^oSee Pogge et al. (2010)

^pSee Dressler et al. (2011). The performance is listed for the two available cameras (F/2 and F/4, respectively), which provide different fields of view and wavelength ranges. One camera can be used at a time

^qIMACS has a cross-dispersed mode that provides the indicated resolution through a 0.6^l slit. See Suttin and McWilliam (2003)

Multi-object fiber spectrographs rely on some means to accurately position a large number of fibers in the telescope focal plane. Accurately machined plug plates (e.g., Shectman 1993), individual mechanisms for positioning each fiber (e.g., Hill and Lesser 1986), and robots with pick-and-place magnetic buttons (e.g., Fabricant et al. 2005a, b) have all been used successfully. Typically, the angle of the fiber must be controlled as well as the position in order to optimize the effective focal ratio for light propagating through the fiber. Fibers are more efficient when fed at faster focal ratios (typically in the range $f/3$ – $f/5$) and it is often necessary to place a micro-optic assembly in front of each fiber in order to convert a relatively slow telescope focal ratio to a more appropriate value.

Collimator design for fiber spectrographs is much simpler than for imaging spectrographs because the fibers can be arranged along a pseudo-slit for input to the spectrograph. The fibers can be arranged at any desired angle along the pseudo-slit in order to control the position of the exit pupil, and at any desired axial position in order to control the field curvature. Camera designs are still demanding because there is still a premium on the camera field of view to provide the maximum number of spectral resolution elements across the largest possible range of wavelength.

When large numbers of fibers are used, they can be sent to more than one spectrograph, as is the case for SDSS (Uomoto et al. 1999; Castander et al. 2001) and 2DF (Lewis et al. 2002), for example. Optical losses and focal ratio degradation lead to efficiency losses in fiber systems, and changes in fiber transmission with time can require special strategies for sky subtraction and relative flux calibration. However, the use of fibers has advantages as well, especially in the flexibility with which objects can be chosen across the available field of view and in the number of objects that can be observed when multiple fibers feed one or more spectrographs.

6 Integral Field Spectrographs

An *integral field spectrograph* divides an area of the sky into spatial resolution elements, and then produces a spectrum for each of these “spaxels.” The result is often thought of as a three-dimensional data cube, with wavelength and the two spatial dimensions varying along the three axes of the cube. Integral field spectrographs have most typically been used to map the velocity field and local star-formation properties across the disks of galaxies.

The simplest way to make an integral field spectrograph is to assemble a densely packed bundle of optical fibers in the telescope focal plane, and then to rearrange the fibers along the pseudo-slit of a fiber spectrograph. Alternatively, an image slicer can be used to divide a field into a series of slit images and then optically reformat the individual slices along a single long slit. A final possibility is to use a lenslet array to create small and separated images of each spatial element (or of the telescope pupil for each spatial element) at the input to an imaging spectrograph and to disperse the images in a direction such that they do not overlap. This is the method used in the instruments TIGER (Bacon et al. 1995) and SAURON (Bacon et al. 2001).

Multiplexing a large number of spatial resolution elements onto the focal plane of a spectrograph places the same premium on a wide field of view as occurs for an imaging spectrograph. In the case of a massive fiber integral field unit (IFU), it is possible to deploy a large number of independent spectrographs, as is being done for the instruments MUSE (Kosmowski et al. 2011; Bacon et al. 2004) and VIRUS (Hill et al. 2006). Infrared IFU spectrographs have proven to be

particularly effective when used with adaptive optics systems (SINFONI at the VLT, Eisenhauer et al. 2003; NIFS at Gemini, McGregor et al. 2002).

7 Near-IR Spectrographs

The basic optical design considerations for near-IR (1–5 μm) spectrographs are similar to those for spectrographs working at optical wavelengths. The most important differences arise from the properties and availability of IR optical materials and the cost and availability of IR detectors.

Optical materials with good transmission beyond about 1.6 μm are limited to certain crystals, a small number of optical glasses (e.g., S-FPL51, S-FTM16, and S-TIM28 from Ohara Corp.), and fused quartz. Crystals that are particularly useful for their optical and structural properties include CaF₂, ZnSe, and Al₂O₃ (sapphire). Of these, sapphire is the strongest and is useful for making vacuum windows. ZnSe has the highest dispersion and is an important material for prisms. Because larger crystals are often more difficult to produce, the diameters of IR-transmitting optical assemblies are quite limited. While IR-transmitting fused quartz (infrasil) can be produced in very large diameters (>1 m), ZnSe is particularly difficult to produce in diameters larger than about 150 mm.

Detector technology has also been slower to progress in the near-IR than in the optical. Only in the last decade have detectors become available in formats as large as $2\text{k} \times 2\text{k} \times 18 \mu\text{m}$ pixels. Previous detectors were small enough that cameras with relatively small focal planes could cover the available detector area. The correspondingly restricted field of view on the sky and limited length of spectral orders limited the overall performance of spectrographs, keeping them smaller and more modest than designs for the optical. The current generation of instruments on 8 m-class telescopes is the first to have large enough detector areas to demand optical designs with comparable performance to optical spectrographs and to fully exploit the available sizes of IR optical materials (see [► Tables 15-2](#) and [► 15-3](#)).

An obvious difference between near-IR and optical spectrographs is the need to provide a vacuum-cryogenic environment in order to keep the thermal background low. This requirement favors smaller and more compact configurations, but the need to suppress the thermal background leads to other complications as well. Near-IR spectrographs must have places in the optical train to effectively baffle thermal stray light from the warm structures in the telescope that precede the spectrograph. There are two reasons why it is better to do this before the light reaches the slit, if possible. One is that the pupil formed by the collimator is likely to be inaccessible because of the presence of a grating (or prism) and is also likely to be degraded by optical aberrations or diffraction effects. The other is that allowing the high thermal background outside of the pupil to enter the spectrograph at all is likely to result in problems with that background being scattered internally. For these reasons, many near-IR spectrographs reimagine the telescope focal plane onto a cold slit and produce a high-quality intermediate pupil inside of the cryogenic environment where a cold stop can remove the thermal background. This is often accomplished by using an Offner relay (Montagnino and Offner 1976), which folds the beam twice between two spherical mirrors, forming a clean pupil on the second mirror and a new image of the telescope focal plane with the same scale as the original. An Offner relay is free of primary aberrations, making it suitable for wide fields, and provides both a convenient location for the cold stop on the second mirror and a convenient image plane for a cold slit.

■ Table 15-3

A representative sample of near-IR, multi-object, imaging spectrographs on 8 m-class telescopes. Column 3 indicates the wavelength range for spectroscopy and imaging. Columns 4 and 7 give the *approximate* peak transmission efficiency for spectroscopy and imaging, respectively. Values include the telescope efficiency unless otherwise noted, and a range indicates values for different bandpasses when available. Columns 5 and 8 list the field of view for spectroscopy and imaging, respectively. Full wavelength coverage may not be available for slits at all locations in the spectroscopic field of view. Column 6 gives the maximum combined slit length possible when multiple slits are used. Column 9 gives the diameter of the collimated beam at the grating. Column 10 gives the resolution obtained using a 1 arcsec slit.

	Tel. (diam)	λ -range (um)	Spec. eff.	Spec. FOV (arcmin)	Slit length (arcsec)	Imaging eff.	Imaging FOV (arcmin)	Pupil diam (mm)	R ($\lambda/\delta\lambda$)
MOSFIRE ^a	Keck (10 m)	0.9–2.45	[0.3]	6.1 × 3	366	[0.3]	6.1 × 6.1	125	2,300
EMIR ^b	GTC (10 m)	0.9–2.5	...	6 × 4	350	...	6 diam.	150	3,000
KMOS ^c	VLT (8 m)	1.0–2.5	[0.3]	7.2 diam.	2.8 ^c	...	6 diam.	50	1,800–4,200 ^c
Flamingos-2 ^d	Gemini (8 m)	0.9–2.5	...	6 × 2	240	...	6 diam.	100	1,150
MOIRCS ^e	Subaru (8 m)	0.8–2.5	0.24	6 × 4	240	0.31	7 × 4	50	650
Lucifer ^f	LBT (8 m)	0.9–2.5	0.2	4 × 3	240	0.33	4 × 4	100	2,400
MMIRS ^g	Magellan (6.5 m)	0.9–2.5	0.15–0.35	7 × 4	240	0.22–0.5	7 × 7	100	1,200

^aSee McLean (2008, 2010), and <http://www2.keck.hawaii.edu/inst/mosfire>. To be commissioned in 2012

^bSee Garzon et al. (2004). To be commissioned in 2013. Efficiency not yet known.

^cSee Rees et al. (2010) and Sharples et al. (2010). To be commissioned in 2012. KMOS has 24 IFUs each with a 2.8 × 2.8 arcsec fields of view. An image slicer on each IFU feeds 0.2 arcsec

slices into three spectrographs. The resolution given is for 0.2 arcsec slices

^dSee Eikenberry et al. (2008) and Raines et al. (2008). Efficiency not known.

^eSee Ichikawa et al. (2006) and Tokoku et al. (2006)

^fSee Seifert et al. (2003), Ageorges et al. (2010), and Seifert et al. (2010), and links at <http://wiki.lbtto.org/twiki/bin/view/PartnerObserving>

^gSee McLeod et al. (2004). <http://www.cfa.harvard.edu/mmit/mmir.html>, and http://www.cfa.harvard.edu/mmit/mmir/Calibration/thru_asbuilt.jpg

An additional difficulty for near-IR multi-object spectrographs is the fact that multiple slit masks are typically required during a single night. Several different cryogenic mechanisms have been developed to move slit masks in and out of the focal plane. These mechanisms must operate with high reliability because the time required to warm up the instrument in order to service them is invariably measured in days if not weeks. By isolating the slit mask cassette with a gate valve and warming up only this portion of the spectrograph, masks can be exchanged on the instrument MMIRS between the end of one night and the beginning of the next, although all of the time available is required (McLeod et al. 2004). Recently, configurable slit mechanisms have been devised that allow on-the-fly configuration of slit masks within the cryogenic environment (e.g., McLean et al. 2008, 2010). Warm fiber feeds have been used through most of the H-band (e.g., APOGEE on SDSS, Wilson et al. 2010; FMOS on Subaru, Kimura et al. 2010), but the thermal background is too high to use fibers at K.

Two advantages for the design of spectrographs working in the near-IR are that mirror coatings tend to have higher reflectivity in the near-IR than at optical wavelengths and refracting materials have smaller dispersion ($dn/d\lambda$). Reflecting collimators and all-reflecting cameras, such as three-mirror anastigmats (TMAs), can therefore achieve high efficiency, especially when coated with gold. When used as spectrograph cameras, TMAs are typically used in an off-axis configuration and can be challenging to align. High-performance, apochromatic refracting cameras can be designed in the near-IR using a smaller number of elements and material types than are typically required at optical wavelengths. The camera for the FIRE spectrograph is one such example (Simcoe et al. 2008). The reduced element count also results in higher throughput and helps to minimize scattered light.

For work on ground-based telescopes, the forest of atmospheric OH-emission lines that exist in the near-IR places a premium on achieving sufficient resolution ($R > \sim 4,000$) in order to resolve the atmospheric lines and allow the detection of features from faint astronomical sources in the darker intervals where the emission is absent (Martini and DePoy 2000). This fact has motivated the development of a number of near-IR echellette spectrographs on 8 m-class telescopes (see [Table 15-1](#)). Echellette designs are particularly suitable for ground-based cryogenic spectrographs: they do not require high cross dispersion, and they target single objects in small fields of view on the sky, allowing them to make use of available materials in limited sizes and to be packaged in compact volumes.

8 Spectrographs for Extremely Large Telescopes (ELTs)

For the next generation of extremely large telescopes (ELTs), the challenge is to design spectrographs with performance comparable to the instruments that have been built for 8 m-class telescopes. Much of the emphasis in the ELT projects is on adaptive optics, where the gain in collecting area over 8 m-class telescopes is compounded by the gain in resolution. For AO instruments working at the diffraction limit, the problem of spectrograph design is relatively straightforward because the smaller size of the entrance aperture (in arcsec) compensates for the larger size of the telescope. For seeing-limited spectrographs (or for AO instruments which do not work at the full diffraction limit), the problem is much more difficult. Some of the spectrograph designs discussed in previous sections lend themselves to being scaled up to a certain extent, but quickly tend to encounter either the fundamental limits imposed by the availability of optical materials or the practical limits of overall instrument complexity and volume.

A factor of 4 increase in pupil diameter would suffice to scale up a seeing-limited instrument from Magellan (6.5 m) to the Giant Magellan Telescope (~25 m diameter), from an 8 m telescope to the Thirty Meter Telescope (TMT, ~30 m diameter), or from Keck (10 m) to the European Extremely Large Telescope (~39 m diameter). Most of the instruments on 8 m-class telescopes have pupil diameters in the range 100–200 mm. A few are smaller, such as LDSS-3 on Magellan (70 mm, Osip et al. 2004), and a few are larger, such as HiRES on Keck (300 mm, Vogt et al. 1994). One strategy is to accept that very large pupil diameters, in the range 600–1,000 mm, will be required. We refer to this strategy below as a “scaled-pupil.” Such large spectrographs cannot be built using the normal range of transmitting optical materials and are restricted to designs that rely primarily on very large reflecting optics, including Schmidt cameras. A second strategy is to increase the pupil diameter to something near the maximum size that can be accommodated using transmitting materials (about 300 mm) and either accept some level of compromise in spectrograph performance or devise some means to work around it. We refer to this strategy below as a “limited pupil.” As is already the case for spectrographs on 8 m-class telescopes, all proposed spectrographs for ELTs achieve higher resolution by using slits or fibers that are smaller than 1 arcsec wide.

Many spectrograph designs have been proposed for the ELTs with various degrees of design development. We do not attempt to include a complete summary of them here. We have focused instead on a few designs that illustrate the range of proposals that have been considered within the general strategies of scaled-pupil or limited-pupil designs.

8.1 Echelles

As an illustration of the scaled-pupil strategy, Vogt (2006, private communication) originally proposed for TMT an extremely large white-pupil spectrograph with separate red and blue arms, called MTHR. The 1.0×3.5 m echelle gratings would require 36-replica mosaics made from 200×400 mm masters. The cross dispersers would require 6-replica mosaics made from 300×400 mm masters. The collimator mirrors would be 3 m in diameter, and the Schmidt corrector lenses 1.3 m in diameter. The instrument would deliver a resolution of 50,000 for a 1 arcsec slit.

Pupil anamorphism and pupil slicing are two ways to mitigate the requirement for very large grating mosaics for high-resolution spectrographs on ELTs. An anamorphic pupil, which can be created for a small field of view by using cylindrical optics, is elliptical instead of round. The spectrograph resolution is always set by the length of the grating mosaic, but an anamorphic pupil allows the mosaic to be made narrower at the expense of the scale of the final spectrum perpendicular to the dispersion. The required cross dispersion increases and the available slit length in arcseconds decreases, but these may be acceptable compromises, especially for single-object spectrographs.

Pupil slicing is most easily achieved by using a lenslet array to image sections of the telescope pupil onto an array of optical fibers. Each fiber effectively carries the light from a smaller aperture telescope, and the fibers can be arranged in a pseudo-slit at the input of a smaller spectrograph. Again, the penalty is in the required cross dispersion and the available slit length since the spectrograph (or an array of spectrographs) must accommodate the fibers from all of the pupil segments, not just one. An alternative method of pupil slicing is to divide the collimator into two or more segments, with small displacements such that the collimator mosaic forms multiple images of the slit or entrance aperture in the direction perpendicular to the echelle dispersion.

Two white-pupil spectrographs, Q-Spec (Barnes and MacQueen 2008) and G-CLEF (Szentgyorgyi et al. 2010; Jaffe et al. 2010), have been proposed for GMT. The four-armed Q-Spec design uses an anamorphic pupil on a $300 \times 1,700$ mm, 4-replica mosaic echelle grating to achieve a resolution of 30,000 for a 1 arcsec slit. A pupil-slicing mode can double the resolution to 60,000. The two-arm G-CLEF design is fiber-fed and uses a $300 \times 1,200$ mm grating to achieve a resolution of 20,000 for a 1.2 arcsec fiber. A pupil-slicing mode in which each of the seven 8.4 m-diameter GMT mirror segments is imaged onto a separate fiber triples the resolution. Both designs have smaller, cross-dispersing, white pupils that enable the use of refracting cameras, optimized over restricted wavelength ranges.

The proposed high-resolution spectrograph for the E-ELT, CODEX, is a single-object, fiber-fed, white-pupil echelle (Pasquini et al. 2008). It uses an anamorphic pupil slicer and a TMA collimator in double-pass with a 415×200 mm pupil and a 1,700 mm, R4 echelle grating to provide resolutions of $\sim 120,000$ in natural seeing conditions. Like G-CLEF, CODEX works over restricted wavelength ranges (between 350 and 720 nm) and includes a dichroic to enable the use of wavelength-optimized VPH gratings followed by compact, refracting cameras. The entire instrument is enclosed in a thermally controlled vacuum vessel, and wavelength calibration will utilize a laser frequency comb (Murphy et al. 2007).

8.2 Wide-Field Optical Spectrographs

As discussed above, wide-field spectrographs are particularly challenging in that the camera fields of view must be much larger than for single-object spectrographs, accommodating both the magnified field of view on the sky and the angular dispersion from the grating. On an ELT, the very large plate scale at the focal plane makes the collimators very large and increases the overall scale of the camera.

Several designs have been proposed for wide-field spectrographs for TMT. Among those, one design explored a four-armed imaging spectrograph with 600 mm-diameter collimated beams using transmitting VPH or surface-relief grating mosaics to achieve a maximum resolution of 5,600 for a 1 arcsec slit (Pazder et al. 2006). This pupil-scaling strategy led to beam sizes that can only be accommodated with reflecting collimators and Schmidt cameras. The overall scale of a four-armed design would be approximately 8 m in diameter and 8 m long. A smaller alternative currently being pursued is an imaging echellette spectrograph with a 300 mm diameter beam called MOBIE (Bernstein and Bigelow 2008; Bigelow and Bernstein 2010). The higher angular dispersion of the echellette grating makes it possible to achieve a resolution of 6,000 for a 1 arcsec slit. The reduced collimated beam diameter permits the use of red- and blue-optimized transmitting cameras with wide angular fields of view. The cross-dispersed formats of the echellette gratings on separate blue and red arms allow full wavelength coverage (300–1,000 nm) in a single exposure.

The baseline optical design for GMT includes a corrector for a 20 arcmin diameter field, which makes it possible to implement a fly's-eye design with chief rays perpendicular to the telescope focal plane as discussed above. The imaging spectrograph concept (Marshall et al. 2012; Jaffe et al. 2010) is a four-armed design with 300 mm diameter collimated beams and VPH gratings, which can achieve resolutions up to about 4,000 with a 1 arcsec slit. The transmitting VPH gratings allow the cameras to be placed very close to the pupil and can provide complete wavelength coverage with a 1-arcsec resolution of 800 and 1,600 on the blue and red

sides, respectively. The GMT field corrector is also suitable for use with a fiber system (MANIFEST, Saunders et al. 2010), and the imaging spectrograph can be readily adapted to fiber input, achieving higher resolution when the fibers are configured as mini-IFUs or image slicers.

Proposals were made for E-ELT instruments that included both fiber-fed and imaging multi-object spectrographs; however, the designs for these were not developed completely, and at this time no wide-field optical spectrograph is planned in the first generation of instruments.

8.3 IR Spectrographs

Several near-IR spectrographs are under development for the first generation of instruments on all three of the ELTs that utilize adaptive optics. In this case, the slits are all smaller than they are for comparable instruments on 8 m-class telescopes. For these instruments, the optical designs are relatively straightforward evolutions from their 8 m-class counterparts, but the smaller slits require greater mechanical complexity and precision, bringing a new array of challenges that we have not discussed in this review. At least one wide-field, seeing-limited near-IR spectrograph has also been proposed (NIRMOS for GMT, Fabricant et al. 2006). Because it is seeing-limited, the design considerations are similar to those for optical instruments with the caveats discussed in [Sect. 6](#) above.

References

- Aderin, M. E. 2004, bHROS installation and system performance. *Proc SPIE*, 5492, 160
- Ageorges, N., Seifert, W., Jütte, M., Knierim, V., Lehmitz, M., Buschkamp, P., & Polsterer, K. 2010, LUCIFER1 commissioning at the LBT. *Proc SPIE*, 7735, 56
- Allington-Smith, J., Graham, M., Content, R., Dodsworth, G., Davies, R., Miller, B. W., Jørgensen, I., Hook, I., Crampton, D., & Murowinski, R. 2002, Integral field spectroscopy with the gemini multiobject spectrograph. I. Design, construction, and testing. *PASP*, 114, 892–912
- Appenzeller, I., Fricke, K., Fürtig, W., et al. 1998, Successful commissioning of FORS1 – the first optical instrument on the VLT. *ESO Messenger*, 94, 1
- Atherton, P. D., Taylor, K., Pike, C. D., Harmer, C.F.W., Parker, N.M., & Hook, R. N. 1982, TAURUS: A wide-field Imaging Fabry-Perot Spectrograph for Astronomy, *MNRAS*, 201, 661
- Bacon, R., Adam, G., Barane, A., Courtes, G., Dubet, D., Dubois, J. P., Emseilem, E., Feruit, P., Georgelin, Y., Monet, G., Pecontai, E., Rousset, A., & Say, F. 1995, 3D Spectroscopy at high spatial resolution and realization of the integral field spectrograph tIGER. *A&AS*, 113, 347
- Bacon, R., Coplin, Y., & Monnet, G. 2001, The SAURON project – I. The panoramic integral-field spectrograph. *MNRAS*, 326, 23
- Bacon, R., Devriendt, J., & Djidel, S., et al. 2004, The second-generation VLT instrument MUSE: science drivers and instrument design. *Proc SPIE*, 5492, 1145
- Baranne, A., Carozzi, N., Comte, G., Courtes, G., Deharveng, J. M., Duflot, R., Monnet, G., & Pellet, A. 1974, Preliminary results of the Baranne white pupil image tube nebular spectrograph. in *Research Programmes for the New Large Telescopes*, Geneva, ed. A. Reiz, 231
- Barden, S. C., Arns, J. A., Colburn, W. S., & Williams, J. 2000, Volume-phase holographic gratings and the efficiency of three simple volume-phase holographic gratings. *PASP*, 112, 809
- Barnes, S., & MacQueen, P. 2008, Q-Spec: a concept for the Giant Magellan Telescope high resolution optical spectrograph. *Proc SPIE*, 7014, 50
- Barnes, S. I., Cottrell, P. L., Albrow, M. D., Frost, N., Graham, G., Kershaw, G., Ritchie, R., Jones, D., Sharples, R., Bramall, D., Schmoll, J., Luck, P., Clark, P., Tyas, L., Buckley, D. A. H., & Brink, J. 2008, The optical design of the Southern African Large Telescope High Resolution Spectrograph: SALT HRS. *Proc SPIE*, 7014, 70140K

- Bernstein, R. A., & Bigelow, B. C. 2008, An Optical Design For A Wide Field Optical Spectrograph for TMT. *Proc SPIE*, 7014, 49
- Bernstein, R. A., Shtetman, S. A., Gunnels, S. M., Mochnacki, S., & Athey, A. E. 2003, MIKE: a double echelle spectrograph for the Magellan telescopes at las campanas observatory. *Proc SPIE*, 4841, 1694
- Bigelow, B. C., & Bernstein, R. A. 2010, Progress of the conceptual design for the MOBIE imaging spectrograph for the Thirty Meter Telescope. *Proc SPIE*, 7735, 74
- Bowen, I. S. 1964, Spectrographs, in *Astronomical Techniques*, ed. W. A. Hiltner (Chicago: University of Chicago Press), 34
- Butler, R. P., Marcy, W., Williams, E., McCarthy, C., & Dosanjh, P. 1996, Attaining doppler precision of 3 m s^{-1} . *PASP*, 108, 500
- Castander, F. J., Nichol, R. C., & Merrelli, A., et al. 2001, The first hour of extragalactic data of the sloan digital sky survey spectroscopic commissioning: the coma cluster. *AJ*, 121, 2331
- Cepa, J., Aguiar, M., & Escalera, V. G., et al. 2000, OSIRIS tunable imager and spectrograph. *Proc SPIE*, 4008, 623
- Crane, J. D., Shtetman, S. A., Butler, R. P., Thompson, C. B., Burley, G. S., & Jones, P. 2010, The Carnegie Planet Finder Spectrograph: integration and testing. *Proc SPIE*, 7735, 773553
- Dekker, H., D'Odorico, S., Kaufer, A., Delabre, B., & Kotzlowski, H. 2000, Design, construction, and performance of UVES. *Proc SPIE*, 4008, 534
- Dressler, A., Bigelow, B., Hare, T., Sutin, B., Thompson, I., Burley, G., Epps, H., Oemler, A., Bagish, A., Birk, C., Clardy, K., Gunnels, S., Kelson, D., Shtetman, S., & Osip, D. 2011, IMACS: The Inamori-Magellan areal camera and spectrograph on Magellan-Baade. *PASP*, 123, 288
- Dumas, D., Fendler, M., Berger, F., Marion, F., Arnaud, A., Vialle, C., Goudon, V., Primot, J., Le Coarer, E., & Ribot, H. 2010, Curved infrared detectors: applications to spectrometry and astronomy. *Proc SPIE*, 7742, 77421V
- Eikenberry, S., Elston, R., Raines, S. N., Julian, J., Hanna, K., Warner, C., Julian, R., Bandyopadhyay, E., Bennett, J., Bessoff, A., Branch, M., Corley, R., Dewitt, C., Eriksen, J., Frommeyer, S., Gonzalez, A., Herlevich, M., Hon, D., Marin-Franch, A., Marti, J., Murphey, C., Rambold, W., Rashkin, D., Leckie, B., Gardhouse, W. B., Fletcher, M., Hardy, T., Dunn, J., & Wooff, R. 2008, FLAMINGOS-2: the facility near-infrared wide-field imager and multi-object spectrograph for Gemini. *Proc SPIE*, 7014, 40
- Eisenhauer, F., Abuter, R., & Bickert, K., et al. 2003, SINFONI – integral field spectroscopy at 50 milli-arcsecond resolution with the ESO VLT. *Proc SPIE*, 4841, 1548
- Epps, H. W. 1990, Camera designs for Keck Observatory LRIS and HIRES spectrometers. *Proc SPIE*, 1235, 550
- Faber, S. M., Phillips, A. C., & Kibrick, R. I., et al. 2003, The DEIMOS spectrograph for the Keck II Telescope: integration and testing. *Proc SPIE*, 4841, 1657
- Fabricant, D. G., Fata, R. G., & Epps, H. W. 1998, Binospec: a dual-beam wide-field optical spectrograph for the converted MMT. *Proc SPIE*, 3355, 232
- Fabricant, D. G., Epps, H. W., Brown, W., Fata, R., & Mueller, M. 2003, The development of Binospec and its optics. *Proc SPIE*, 4841, 1134
- Fabricant, D., Fata, R., Roll, J., Hertz, E., Caldwell, N., Gauron, T., Geary, J., McLeod, B., Szentgyorgyi, A., Zajac, J., Kurtz, M., Barberis, J., Bergner, H., Brown, W., Conroy, M., Eng, R., Geller, M., Goddard, R., Honsa, M., Mueller, M., Mink, D., Ordway, M., Tokarz, S., Woods, D., Wyatt, W., Epps, H., & Dell'Antonio, I. 2005a, Hectospec, the MMT's 300 optical fiber-fed spectrograph. *PASP*, 117, 1411
- Fabricant, D. G., Fata, R. G., & Roll, J., et al. 2005b, Hectospec, the MMT's 300 optical fiber-fed spectrograph. *PASP*, 117, 838
- Fabricant, D., Hertz, E., Brown, W., McLeod, B., Angel, R., & Lloyd-Hart, M. 2006, A wide-field IR spectrograph for the Giant Magellan Telescope. *Proc SPIE*, 6269, 64
- Garzon, F., Abreu, D., Barrera, S., Correa, S., Diaz, J. J., Fragoso, A. B., Fuentes, F. J., Gago, F., Gonzalez, C., Lopez, P., Manescau, A., Patron, J., Perez, J., Redondo, P., Restrepo, P., Sanchez, V., & Villegas, V. 2004, EMIR: the GTC NIR multi-object imager-spectrograph. *Proc SPIE*, 5492, 1187
- Hammersley, P., Christensen, L., Dekker, H., Izzo, C., Selman, F., Bristow, P., Bourget, P., Castillo, R., Downing, M., Haddad, N., Hilker, M., Lizon, J.-L., Lucuix, C., Mainieri, V., Mieske, S., Reinero, C., Rejkuba, M., Rojas, C., Smette, A., Urrutia Del Rio, J., Valenzuela, J., & Wolff, B. 2010, Upgrading VIMOS. *Messenger*, 142, 8
- Hartmann, P., & Jedamzik, R. 2006, Large optical glass lenses for ELTs. *Proc SPIE*, 6273, 62730H
- Hill, J. M., & Lesser, M. P. 1986, Deployment of the MX spectrometer. *Proc SPIE*, 627, 303
- Hill, G. J., MacQueen, P. J., & Phillip, J., et al. 2006, VIRUS: a massively replicated integral-field spectrograph for HET. *Proc SPIE*, 6269, 79
- Hinkle, K. H., Cuberly, R. W., Gaughan, N. A., Heynssens, J. B., Joyce, R. R., Ridgway, S. T., Schmitt, P., & Simmons, J. E. 1998, Phoenix: a

- cryogenic high-resolution 1- to 5- μm infrared spectrograph. *Proc SPIE*, 3354, 810
- Hinkle, K. H., Joyce, R. R., Sharp, N., & Valenti, J. A. 2000, Phoenix: operation and performance of a cryogenic high-resolution 1–5 μm infrared spectrograph. *Proc SPIE*, 4008, 720
- Hinkle, K. H., Blum, R., Joyce, R. R., Ridgway, S. T., Rodgers, B., Sharp, N., Smith, V., Valenti, J., & van der Blik, N. 2002, The Phoenix spectrograph at Gemini South. *Proc SPIE*, 4834, 353
- Hook, I., Jørgensen, I., Allington-Smith, J. R., Davies, R. L., Metcalfe, N., Murowinski, R. G., & Cramp-ton, D. 2004, The Gemini-North multiobject spectrograph: performance in imaging, long-slit, and multi-object spectroscopic modes. *PASP*, 116, 425
- Ichikawa, T., Suzuki, R., Tokoku, C., Katsuno, Y., Uchimoto, K., Konishi, M., Yoshikawa, T., Yamada, T., Tanaka, I., Omata, K., & Nishimura, T. 2006, MOIRCS: multi-object infrared camera and spectrograph for SUBARU. *Proc SPIE*, 6269, 16
- Jaffe, D., Mar, D. J., Warren, D., & Segura, P. R. 2006, GMTNIRS: the high resolution near-IR spectrograph for the Giant Magellan Telescope. *Proc SPIE*, 6269, 143
- Jaffe, D. T., Wang, W., Marsh, J. P., Deen, C. P., Kelly, D., & Greene, T. P. 2008, Fabrication and test of silicon grisms for JWST-NIRCam. *Proc SPIE*, 7010, 104
- Jaffe, D., Fabricant, D., & Hinz, P., et al. 2010, Science instrument development for the Giant Magellan Telescope. *Proc SPIE*, 7735, 72
- Jedamzik, R., & Hartmann, P. 2006, Large optical glass blanks for astronomy. *Proc SPIE*, 5494, 382
- Kashikawa, N., et al. 2002, FOCAS: the faint object camera and spectrograph for the Subaru Telescope. *PASJ*, 54, 819
- Kimura, M., Maihara, T., & Iwamuro, F., et al. 2010, The fiber multi-object spectrograph (FMOS) for the Subaru Telescope. *PASJ*, 62, 1135
- Kosmalski, J., Parés, L., & Seifert, W. et al. 2011, Optical design of the VLT/MUSE instrument. *Proc SPIE*, 8167, 816716
- Le Fèvre, O., Saisse, M., Mancini, D., Brau-Nogue, S., Caputi, O., Castinel, L., D'Odorico, S., Garilli, B., Kissler-Patig, M., Lucuix, C., Mancini, G., Pauget, A., Sciarretta, G., Scodreggio, M., Tresse, L., & Vettolani, G. 2003, Commissioning and performances of the VLT-VIMOS instrument. *Proc SPIE*, 4841, 1670
- Lewis, I. J., Cannon, R. D., & Taylor, K., et al. 2002, The Anglo-Australian observatory 2dF facility. *MNRAS*, 333, 279
- Marcy, G., & Butler, R. P. 1992, Precision radial velocities with an iodine absorption cell. *PASP*, 104, 270
- Marshall, J. L., Bures, S., Thompson, I. B., Shectman, S. A., Bigelow, B. C., Burley, G., Birk, C., Estrada, J., Jones, P., Smith, M., Kowal, V., Castillo, J., Storts, R., & Ortiz, G. 2008, The MagE spectrograph. *Proc SPIE*, 7014, 169
- Marshall, J. L., DePoy, D. L., & Shectman, S. A., et al. 2012, The GMACS spectrograph for the GMT, in American Astronomical Society Meeting Abstracts, AAS Meeting, 219, #422.13
- Martini, P., & DePoy, D. L. 2000, Optimal resolutions for IR spectroscopy through the OH airglow. *Proc SPIE*, 4008, 695
- Mayor, M., Pepe, F., & Queloz, D., et al. 2003, Setting new standards with HARPS. *The Messenger*, 114, 20
- McCarthy, J. K., Cohen, J. G., Butcher, B., Cromer, J., Croner, E., Douglas, W. R., Goeden, R. M., Grewal, T., Lu, B., Petrie, H. L., Weng, T., Weber, B., Koch, D. G., & Rodgers, J. M. 1998, Blue channel of the Keck low-resolution imaging spectrometer. *Proc SPIE*, 3355, 81
- McGregor, P., Hart, J., Conroy, P., Pfitzner, L., Bloxham, G., Jones, D., Downing, M., Dawson, M., Young, P., Jarnyk, M., & van Harmelen, J. 2002, Gemini near-infrared integral field spectrograph (NIFS). *Proc SPIE*, 4841, 178
- McLean, I. S., Becklin, E. E., & Bendiksen, O., et al. 1998, Design and development of NIRSPEC: a near-infrared echelle spectrograph for the Keck II telescope. *Proc SPIE*, 3354, 566
- McLean, I. S., Steidel, C. C., Matthews, K., Epps, H., & Adkins, S. 2008, MOSFIRE: a multi-object near-infrared spectrograph and imager for the Keck Observatory. *Proc SPIE*, 7014, 99
- McLean, I. S., Steidel, C. C., Epps, H., Matthews, K., Adkins, S., Konidaris, N., Weber, B., Aliado, T., Brims, G., Canfield, J., Cromer, J., Fucik, J., Kulas, K., Mace, G., Magnone, K., Rodriguez, H., Wang, E., & Weiss, J. 2010, Design and development of MOSFIRE: the multi-object spectrometer for infrared exploration at the Keck Observatory. *Proc SPIE*, 7735, 47
- McLeod, B. A., Fabricant, D., Geary, J., Martini, P., Nystrom, G., Elston, R., Eikenberry, S. S., & Epps, H. 2004, MMT and Magellan infrared spectrograph. *Proc SPIE*, 5492, 1306
- Montagnino, L., & Offner, A. 1976, Design and testing with a reflective null system. *NASA Spec Publ*, 392, 135
- Murphy, M. T., Udem, Th., & Holzwarth, R., et al. 2007, High-precision wavelength calibration of astronomical spectrographs with laser frequency combs. *MNRAS*, 380, 839

- Oke, J. B., Cohen, J. G., Carr, M., Cromer, J., Din-gizian, A., & Harris, F. H. 1995, The Keck low-resolution imaging spectrometer (LRIS). *PASP*, 107, 375
- Osip, D., Phillips, M., & Bernstein, R. A., et al. 2004, First-generation instruments for the Magellan telescopes: characteristics, operation, and performance. *Proc SPIE*, 5492, 49
- Pasquini, L., Avila, G., Blecha, A., Cacciari, C., Cayatte, V., Colless, M., Damiani, F., de Propris, R., Dekker, H., di Marcantonio, P., Farrell, T., Gillingham, P., Guinouard, I., Hammer, F., Kaufer, A., Hill, V., Marteau, M., Modigliani, A., Mulas, G., North, P., Popovic, D., Rossetti, E., Royer, F., Santin, P., Schmutzer, R., Simond, G., Vola, P., Waller, L., & Zoccali, M. 2002, Installation and commissioning of FLAMES, the VLT multifibre facility. *The Messenger*, 110, 1
- Pasquini, L., Avila, G., & Dekker, B., et al. 2008, CODEX: the high-resolution visual spectrograph for the E-ELT. *Proc SPIE*, 7014E, 51
- Pazder, J. S., Fletcher, M., & Morbey, C. 2006, The optical design of the wide field optical spectrograph for the Thirty Meter Telescope. *Proc SPIE*, 6269, 97
- Pogge, R. W., Atwood, B., Brewer, D. F., Byard, P. L., Derwent, M. A., Gonzalez, R., Martini, P., Mason, J. A., O'Brien, T. P., Osmer, P. S., Pappalardo, D. P., Steinbrecher, D. P., Teiga E. J., & Zhelem, R. 2010, The multi-object double spectrographs for the Large Binocular Telescope. *Proc SPIE*, 7735, 77350A
- Perruchot, S., Arnold, L., & Bouchy, F., et al. 2011, Higher-precision radial velocity measurements with the SOPHIE spectrograph using octagonal-section fibers. *Proc SPIE*, 8151, 815115.P
- Raines, S. N., Eikenberry, S., Bandyopadhyay, R., Julian, J., Hanna, K., Warner, C., Julian, R., Bennett, J. G., DeWitt, C. N., Frommeyer, S., Gonzalez, A., Herlevich, M., & Murphey, C. 2008, Characterization and testing of FLAMINGOS-2: the Gemini facility near-infrared multi-object spectrometer and wide-field imager. *Proc SPIE*, 7014, 70143
- Rees, P., Cirasuolo, M., Lewis, I. J., & Todd, S. P. 2010, First end-end performance testing and results for KMOS. *Proc SPIE*, 7735, 77351
- Saunders, W., Colles, M., & Saunders, I., et al. 2010, MANIFEST: a many-instrument fiber-positioning system for GMT. *Proc SPIE*, 7735, 205
- Schroeder, D. J. 2000, *Astronomical Optics* (London: Academic)
- Seifert, W., Appenzeller, I., Baumeister, H., Bizenberger, P., Bomans, D., Dettmar, R., Grimm, B., Herbst, T., Hofmann, R., Juette, M., Laun, W., Lehmitz, M., Lemke, R., Lenzen, R., Mandel, H., Polsterer, K., Rohloff, R., Schuetz, A., Seltmann, A., Thatte, N., Weiser, P., & Xu, W. 2003, LUCIFER: A multi-mode NIR instrument for the LBT. *Proc SPIE*, 4841, 962
- Seifert, W., Ageorges, N., Lehmitz, M., Buschkamp, P., Knierim, V., Polsterer, K., & Germeroth, A. 2010, Results of LUCIFER1 commissioning. *Proc SPIE*, 7735, 292
- Sharples, R., Bender, R., & Berbel, A. A., et al. 2010, Recent progress on the KMOS multi-object integral-field spectrograph for ESO VLT. *Proc SPIE*, 7735, 15
- Shectman, S. A. 1993, Fiber-optic spectroscopy at the las campanas 2.5-meter telescope, in *ASP Conf. Ser. 37, Fiber-Optics in Astronomy II*, ed. P. M. Gray (San Francisco, CA: ASP), 26
- Shectman, S. A. 1994, The optical design of the Magellan project 6.5-meter telescope. *Proc SPIE*, 2199, 558
- Sheinis, A. I., Bolte, M., Epps, H. W., Kibrick, R. I., Miller, J. S., Radovan, M. V., Bigelow, B. C., & Sutin, B. M. 2002, ESI, a New Keck observatory echellette spectrograph and imager. *PASP*, 114, 851
- Simcoe, R. A., Burgasser, A. J., Bochanski, J. J., Schechter, P. L., Fishner, J., Burgasser, A. J., Bernstein, R. A., Bingleow, B. C., Pipher, J. L., Forrest, W., McMurtry, C., & Smith, M. J. 2010, The FIRE infrared spectrometer at Magellan: construction and commissioning. *Proc SPIE*, 7735, 773538
- Simcoe, R. A., Burgasser, A. J., Fishner, J., Schechter, P. L., Smith, M., Bernstein, R. A., Bingleow, B. C., Forrest, W., McMurtry, C., & Pipher, J. L. 2008, FIRE: a near-infrared cross-dispersed echellette spectrometer for the Magellan telescopes. *Proc SPIE*, 7014, 701427
- Spanò, P., Delabre, B., & Norup Sørensen, A., et al. 2006, The optical design of X-Shooter for the VLT. *SPIE Conf Ser.* 6269, 62692
- Suttin, B., & McWilliam, A. 2003, Multi-object high-resolution echellette spectroscopy with IMACS. *Proc SPIE*, 4841, 1357
- Szentgyorgyi, A., Furesz, G., Cheimets, P., Conroy, M., Eng, R., Fabricant, D., Fata, R., Gauron, T., Geary, J., McLeod, B., Zajac, J., Amato, S., Bergner, H., Caldwell, N., Dupree, A., Goddard, R., Johnston, E., Meibom, S., Mink, D., Pieri, M., Roll, J., Tokarz, S., Wyatt, W., Epps, H., Hartmann, L., & Meszaros, S. 2011a, HectoChelle: a multiobject optical echelle spectrograph for the MMT. *PASP*, 123, 1188
- Szentgyorgyi, A., Furesz, G., & Frebel, A., et al. 2011, The GMT-CFA-Carnegie-Catolica Large Earth

- Finder (G-CLEF): A Fiber-fed, optical echelle spectrograph for the Giant Magellan telescope. *Bull Am Astron Soc*, 43, #413.07
- Terada, H, Yuji, I., Kobayashi, N, Yasui, C., Pyo, T, Usuda, T, Hayashi, M, & Kawakita, H. 2008, High Resolution Spectrograph Unit (HRU) for the SUBARU/IRCS. *SPIE*, 714, 103
- Tokoku, C., Suzuki, R., Omata, K., Konishi, M., Yoshikawa, T., Akiyama, M., Tanaka, I., Ichikawa, T., & Nishimura, T. 2006, Infrared multi-object spectrograph of MOIRCS. *Proc SPIE*, 6269, 4
- Udry, S., Mayor, M., & Benz, W. 2006, The HARPS search for southern extra-solar planets. *A&A*, 447, 261
- Uomoto, A., Smee, S., & Rockosi, C., et al. 1999, The sloan digital sky survey spectrographs. *Bull Am Astron Soc*, 31, 1501
- Veilleux, S., Weiner, B. J., & Rupke, D. S. N., et al. 2010, MMTF: the Maryland-Magellan tunable filter. *AJ*, 139, 145
- Vernet, J., Dekker, H., & D'Odorico, S., et al. 2011, X-shooter, the new wide band intermediate resolution spectrograph at the ESO VLT. *A&A*, 536, A105
- Vogt, S. S., Allen, S. L., Bigelow, B. C., Bresee, L., Brown, B., Cantrall, T., Conrad, A., Couture, M., Delaney, C., Epps, H. W., Hilyard, D., Hilyard, D. F., Horn, E., Jern, N., Kanto, D., Keane, M. J., Kibrick, R. I., Lewis, J. W., Osborne, J., Pardeilhan, G. H., Pfister, T., Ricketts, T., Robinson, L. B., Stover, R. J., Tucker, D., Ward, J., & Wei, M. Z. 1994, HIRES: the high resolution echelle spectrometer on the Keck 10-m telescope. *Proc SPIE*, 2198, 362
- Walker, G. M., Mateo, M., Olszewski, E. W., Bernstein, R. A., Sen, B., & Woodrooffe, M. 2006, The Michigan/MIKE fiber system of stellar radial velocities in dwarf spheroidal galaxies: acquisition and reduction of data. *AJSS*, 171, 389
- Wilson, J. C., Hearty, F., & Skrutski, M. F., et al. 2010, The apache point observatory galactic evolution experiment (APOGEE) high-resolution near-infrared multi-object fiber spectrograph. *Proc SPIE*, 7735, 46

Index

A

Aberrations, 299–301
Acquisition, 6, 29–32
Active control matrix, 113–120
Active control systems, 102
Active cooling, 386, 426
Active optics, 33–36, 154, 156–158, 160, 163, 165, 167, 168, 171–173, 187, 188, 210, 212–225, 228–231, 233–235, 238
Actuators, 102, 105, 108, 109, 112–120, 129–132
Adaptive optics (AO), 25, 26, 35–38, 142, 143, 155, 167, 168, 170, 172, 174–177, 187–190, 231, 234, 236, 238
Airy pattern (disk), 19
Alignment, 187, 197–200, 212, 217, 218, 221, 226, 227, 229, 235, 300, 309–310
Alignment of secondary mirror, 127–128
All sky monitor, 52, 56–59, 85
Altitude-azimuth system, 30
Anamorphic magnification, 597
Angular field of view, 11
Anisotropy, 432–435, 442, 443, 445, 449–451, 457–459, 465–471, 476
Antenna, 319, 321, 324–338, 345, 347, 349, 351, 353, 354, 357
 cassegrain, 345
 G/T value, 329
 Hertz dipole, 325
 nasmyth, 328, 329
 reciprocity, 325
AO. *See* Adaptive optics (AO)
Aperture
 effective, 325
 efficiency, 326
 geometric, 326–328
Aperture stop, 10, 11, 13, 17
Aperture synthesis, 27, 28
Aplanatic systems, 15
Arcade, 449
Areal density, 409, 410, 413
Artificial wavelength, 124
Asphericity, 102, 106–108
Astatic, support, 203–204
Astronomical detectors for, 509
Astronomical instrumentation, 508, 509, 538
Atmosphere, 187, 190, 191, 197–198, 218, 220, 227, 232–234, 236
 refraction, 263
 seeing, 243, 248, 249, 268, 269, 275
 turbulence, 247, 256

Attitude, 30, 31
Auto-guiding, 32
Autonomous observatory, 47

B

Background, 366–369, 371–376, 378–380, 383, 384, 390, 396, 399, 414, 420, 422, 423, 426
Back-illumination, 547, 557
Basic theory, 243–246, 260
Beam
 combination, 176
 efficiency, 326
 solid angle, 326
 width, 327
Black body radiation, 322–323
Blockage, 297, 298, 300, 302
Bolometers, 566–572, 581
Bootstrapping, 255, 278
Borosilicate glass, 140, 141, 145, 150, 152
Brightness temperature, 323, 324, 327, 328, 332

C

Cameras, 508–513, 515, 517, 519, 520, 522, 531, 533, 535, 589–591, 594–596, 598, 599, 602–604, 607–609, 611–613
Cassegrain designs, 299, 301
Catadioptric cameras, 594
Charge-coupled device (CCD), 543–562
 camera, 51–53, 57, 59, 60, 62, 63, 65–67, 80
Charge transfer coefficient (CTC), 20
Cherenkov radiation, 492
Chopping, 300–302
Chromatic aberration, 13, 14, 22
Closure phase, 249–251, 263, 264, 268–271
CMOS sensor, 543, 544, 552, 555–558, 562
COBE. *See* COsmic Background Explorer (COBE)
Coherence time, 24, 26
Coherent, 18
Cold dark matter, 436
Collimator, 589–591, 594, 596–598, 601–605, 608, 609, 611–613
Comatic aberration (coma), 15, 22
Commercial off the shelf (COTS) components, 47, 57, 60
Communication protocol, 45, 85–87
Compton scattering, 484, 485, 488
Contrast, 13, 20–24, 28, 29, 37, 38

Contrast transfer function (CTF), 20, 23, 24, 28

Control

closed loop, 213–215

open loop, 213–215

Coronagraphs, 511–512

COSMIC Background Explorer (COBE), 435, 437, 446, 467–472, 476

Cosmic microwave background radiation (CMB), 431–473

CTF. *See* Contrast transfer function (CTF)

Curvature sensing, 125

D

Delay line, 249, 251, 254, 255, 260–267, 270, 271, 275, 276

Delivered image quality, 6, 14, 24, 26, 29, 32, 33, 40

Detectors, 507–538

Differential microwave radiometer (DMR), 470–472

Diffraction effects, 257

Diffraction limit, 142–143, 156, 168, 170, 176, 177, 188, 190, 193, 215, 234, 236–238

Dioptic cameras, 594–596

Dispersion, 9

compensation, 261

DMR. *See* Differential microwave radiometer (DMR)

Drive system, 30, 32

E

Echelle, 588, 592, 594, 596–602, 611–613

Edge sensors, 102, 109, 112–113, 115, 117, 120, 129–132, 306, 307

E-ELT. *See* European extremely large telescope (E-ELT)

Electromagnetic radiation (EMR), 2–10, 14, 17, 18, 21, 22, 28, 33, 38, 40

Enclosure, 6, 7, 11, 35, 38–39

Entrance pupil, 10, 11, 13, 23

Equatorial system, 30

Error budget, 29

Error multipliers for control systems, 113, 117

Error sources

misalignment, 188–190

polishing, 189

temperature, 190

wind, 190

Etendue, 12, 49, 50, 72, 80, 81

European extremely large telescope (E-ELT), 104, 130, 131, 133

Event broker, 87

Exposure time, 373–379, 391, 426

Extensive air shower, 487–490, 494, 496

Extinction, 366–367, 426

F

Fairing, 402–404

Far infrared absolute spectrophotometer (FIRAS), 435, 470, 476

Field aberrations, 192–194, 197, 213, 219, 237

Field of view (FOV), 4, 10, 11, 15, 16, 22, 24, 29, 32, 37, 38

Field stop, 10, 11

Filter, 190, 195, 202, 215, 218, 233, 234

FIRAS. *See* Far infrared absolute spectrophotometer (FIRAS)

Flux

density, 322, 324, 332

total, 324

Focal length, 10–14, 16, 20, 21

Focal plane array, 54, 71, 80, 82

Focal ratio, 11, 22

Focus mode, 112, 118, 120

FOV. *See* Field of view (FOV)

Fraunhofer diffraction, 18

Fried's parameter (r_0), 23, 24, 27

Fringe tracking, 243, 255, 275–278

Full width at half maximum (FWHM), 19, 24

Future telescopes, 140, 143, 178–179

FWHM. *See* Full width at half maximum (FWHM)

G

Galactic emission, 433

Gamma ray, 481–504

Gas fusion, 141, 166

GCN. *See* GRB Coordinates Network (GCN)

Giant Magellan telescope (GMT), 104, 130–132, 139, 140, 148, 154, 158–160, 162–165, 178–179

Giant segmented-mirror telescope, 101, 104, 111, 127, 130–133

GLAO. *See* Ground layer AO (GLAO)

GMT. *See* Giant Magellan telescope (GMT)

Gran Telescopio Canarias (GTC), 103, 127, 128, 131

Gratings, 588–594, 596–606, 609, 610, 612, 613

Gravitational deflection, 300, 303–305

Grazing incidence, 3, 17, 18

GRB Coordinates Network (GCN), 65, 66, 68, 76, 77, 86, 88

GRB follow-up, 56, 62, 67, 68, 76

Gregory designs, 296, 299–302

Gregory system, 328, 329

Ground layer AO (GLAO), 37, 38

GTC. *See* Gran Telescopio Canarias (GTC)

Gyroscope, 384, 414–417

H

HET. *See* Hobby-Eberly telescope (HET)

History of segmented-mirror telescopes, 102–104

Hobby-Eberly telescope (HET), 103, 117, 119,
128–129
Honeycomb sandwich, 140, 143, 146, 151–154
Hubble space telescope (HST), 7, 22, 30, 369, 370,
376–378, 381–385, 388–390, 392,
394–398, 401, 402, 405–407, 409–411,
413–416, 418, 420, 422–424, 426, 427
Huygens-Fresnel, 18

I

Image quality, 2, 4–6, 14–16, 24–27, 29, 32, 33, 35,
38, 40, 187–192, 200, 210, 212, 215, 218,
219, 224, 226, 227, 229, 234, 236
Imaging atmospheric Cherenkov telescope, 494–503
Infrared detectors, 565–584
Instruments, 6, 7, 16, 22, 26, 33, 36–38
Interferometer, 321, 322, 330–332, 334, 335, 337,
338, 340, 347, 351, 356
noise, 330, 332, 508, 509, 521–523
Interferometry, 243, 245, 248, 249, 251, 253, 261,
270–273, 275, 278
Ion beam figuring, 108
Isoplanatic angle, 24
Isoplanatic patch, 24, 36, 37

J

James Webb space telescope (JWST), 101, 125, 128,
130, 367–371, 377, 378, 381, 382,
388–396, 399–402, 404, 405, 407–414,
416–420, 425–427

K

Keck active control system, 114, 117
Keck telescopes, 101, 102, 105, 107, 111, 112, 117,
118, 120, 121, 125, 126, 128
diffraction patterns, 111

L

Large area multi-object spectrographic telescope
(LAMOST), 129
Large binocular telescope (LBT), 140, 143, 145–147,
149, 153, 154, 157–163, 167–180
Laser guide star AO (LGS-AO), 37, 38
Launch vehicle, 381, 390, 392, 393, 396, 397,
400–403, 408, 410, 423
LBT. *See* Large binocular telescope (LBT)
Lensing, 458
LGS-AO. *See* Laser guide star AO (LGS-AO)
Lightweight, 384, 388, 400, 404, 408–410, 412,
419, 427
Lightweight optics, 140, 143, 169
Lookup tables, 34

M

Mach-Zehnder interferometer, 125
Magnetic rheological figuring (MRF), 108
Magnification, 10–13, 16, 22
Main beam, 324, 326, 327, 331, 342
MCAO. *See* Multi-conjugate AO (MCAO)
Metrics for telescope quality, 188
Microwave kinetic induction detectors (MKID),
582–584

Mirror

cells, 34
control, 306–307
meniscus, 204, 206, 212–224, 235
seeing, 141, 143, 173
segmented, 197, 198, 200, 221, 224, 236
structured, 197, 201, 202, 206, 235
support, 144, 155–157, 167, 171–174
force-based, 203–204
position-based, 202–203, 223, 224
MKID. *See* Microwave kinetic induction detectors
(MKID)
MMT. *See* Multiple mirror telescope (MMT)
Modes, elastic, 194–200, 202, 210–212, 220–222,
224, 226, 227, 229–234
Modulation transfer function (MTF), 20
Monolithic mirror, 101, 104, 132
Multi-conjugate AO (MCAO), 37, 38
Multiple mirror telescope (MMT), 103, 104, 140,
154, 162, 167–168, 173

N

Natural frequency, 305, 306, 308
Noll coefficients, 27
Nyquist theorem, 323

O

Observatory, 2, 6–7, 17, 37, 39
Off-axis, 297, 298, 300–302
On-axis images, 10, 11, 22
OPD. *See* Optical path difference (OPD)
Open loop tracking, 32, 35
Optical alignment, 102, 121–126, 128, 132
Optical axis, 10, 12–15, 21, 22, 24–26, 31, 34, 35
Optical delay, 260–262, 265
Optical path, 9, 10, 14
Optical path difference (OPD), 25, 28
Optical telescope assembly (OTA), 381–384, 389,
405–407, 409, 419
Optical testing, 158, 164, 165, 172
Optics, 296–302, 309, 310
Orbit, 381–383, 385, 388, 390, 392–394, 396–403,
405, 406, 408, 410–413, 416, 422–427
OTA. *See* Optical telescope assembly (OTA)

P

Pair production, 483–488, 502
 Paraxial optical system, 10
 Paraxial rays, 10–12, 21
 Passive cooling, 390, 400, 423, 426
 Passive telescopes, features, 188, 200, 210–212
 Performance, telescope, 191, 210, 212
 Phase contrast interferometry, 310
 Phased array imaging, 168, 176–177
 Phasing, 103, 104, 121–126, 129–131
 Photoconductors, 566, 571–575, 577–581
 PIXIE, 439, 476
 Planck, 435, 437–441, 445, 447, 451, 454, 467, 468, 471, 474–476
 Planck function, 323, 349
 Plate scale, 12, 20, 300–301
 Pointing and control, 415, 419
 Pointing error, 303, 305, 308, 309
 Pointing model, 31, 32
 Point spread function (PSF), 20, 21, 23, 24, 25, 28, 32
 Polarimeters, 508, 509, 520–521
 Polarization, 256, 257, 275, 432, 433, 438–443, 445, 447, 449–452, 454–459, 461–465, 467, 468, 470, 471, 473, 474, 476
 Polishing, 141, 142, 153–155, 159–161, 163, 167, 172, 174, 178
 Power normalized pattern, 326, 327
 Primary mirror, 140–145, 148, 150, 154, 156–163, 166–175, 178, 179, 369, 381, 383, 385, 388–395, 400–405, 407–414, 418, 422, 426, 427
 Prime focus, 15–17, 33, 169, 170, 172
 Prisms, 591–594, 597, 599–602, 609
 Pseudo-inverse matrix, 116
 PSF. *See* Point spread function (PSF)
 Pupil, 589–591, 596–600, 602–604, 606, 608–610, 612, 613
 magnification, 590, 591, 598
 Pyramid sensing, 125

Q

Quantum efficiency, 543–545, 547, 549, 551, 557, 558

R

Radius of curvature, 11, 15, 16
 Rapid follow-up, 46, 51, 54, 74, 86, 94
 Rayleigh-Jeans law, 323
 Rays, 7, 8, 10, 12–15, 17, 18, 21, 23
 Reaction wheel, 382, 384, 388, 390, 414, 416, 418, 419
 Readout noise, 543, 544, 549–552, 555, 558, 562

Reflectivity, 8, 9, 17
 Reflectors, 14, 29
 Refractors, 3, 12–14
 Remote telescope markup language (RTML), 86, 94
 Remote telescope system (RTS), 47, 56, 86
 Resolution angular, 321, 324, 326–328, 330, 334, 335, 339, 343, 345, 347, 351
 Robotic telescope, 45–48, 51, 52, 54, 64, 66, 68, 73–75, 83, 85, 86, 90–92, 94
 RTML. *See* Remote telescope markup language (RTML)

S

Scaling, 195, 206, 222
 Scanning, 300–303, 305
 SCAO. *See* Single conjugate AO (SCAO)
 Schwarzschild (conic) constant, 15
 Secondary support, 297, 300, 302
 Seeing, 187, 188, 190, 191, 193, 201, 203, 210, 212, 215, 220, 228, 230–232, 234–236, 238
 Seeing-limited, 187, 188, 191, 193, 201, 203, 210, 212, 215, 228, 232, 234, 238
 Segment
 gaps, 109, 110
 polishing, 102, 108
 support, 102, 108–109
 Seidel aberrations, 21, 22, 26, 34
 Sensitivity, 369–374, 379, 389, 391, 395, 417
 telescope, 319–321
 Servomechanisms, 30, 33
 Shack-Hartmann phasing, 122–125
 Shearing interferometer, 310
 Sidereal rate, 30
 Signal-to-noise ratio (S/N), 366, 371–375, 377, 378, 380, 390, 391, 400
 Silicon image sensor, 541–563
 Single conjugate AO (SCAO), 36–38
 Singular value decomposition, 116–118, 120
 Skylert, 87, 88, 90, 94
 S/N. *See* Signal-to-noise ratio (S/N)
 Solar array, 384, 390, 402, 414, 419, 422, 425
 Southern African Large Telescope (SALT), 103, 128–129
 Spacecraft, 382, 383, 387, 390, 399, 402, 403, 415–419, 424, 425
 Space telescopes, 361–427
 Spatial filtering, 270, 272, 273, 275, 278
 Spatial frequency, 20–24, 27, 28, 33, 36
 Spatial interferometers, 3
 Speckle imaging, 25
 Spectrometers, 508–510, 512–522, 535
 Spectrum distortion, 474
 Spherical aberration, 14, 21, 22, 26

Spin casting, 141, 148
 Spitzer space telescope (SST), 369, 381, 382,
 385–388, 390–396, 398, 399, 401–403,
 405, 406, 409, 411, 413, 416, 418, 420,
 423–426
 Standard model, 433
 Steward Observatory Mirror Laboratory, 141, 150
 Stiffness
 modes and mirrors, 195, 197, 201, 202, 211,
 212, 222
 of structures, 187
 Strehl ratio (S), 25, 175, 176, 258, 259
 Stressed lap, 159, 160
 Stressed mirror polishing, 108, 126
 Structure-telescope-mirror cell, 221
 Submillimeter telescopes, 283–310
 Sunshield, 383, 390, 394, 395, 403, 404, 407, 414,
 419, 420, 425, 426
 Sunyaev-Zel'dovich, 459
 Survey telescope, 45, 47, 49, 50, 52, 54,
 55, 94
 System noise, 329

T

Telescope, 2–40, 137–179
 mount, 6, 29–33, 54, 63, 303, 307–309
 fast slewing, 51, 63, 69
 network, 45, 46, 55, 63, 79, 83–85, 90
 structure, 29
 Telescope control system (TCS), 6
 Temperature
 antenna, 324, 327, 328
 brightness, 323, 324, 327,
 328, 332
 fluctuations, 332
 noise, 323, 329
 Temporal fringe encoding, 267, 270
 Thermal deformations, 302–304, 306,
 308, 309

Thirty meter telescope (TMT), 104, 115, 117,
 130–133
 Tracking, 6, 25, 29, 30, 32, 35

U

UV plane, 28, 245

V

Vignetting, 10
 Visibility, 20, 28
 calibration, 249–251
 function, 244, 245, 250
 VOEvent, 86, 87, 94

W

Warping harnesses, 109, 126–127
 Water Cherenkov telescope (WCT), 490–494, 501
 Wavefront
 error, 303, 310
 sensor, 187, 200, 215–220, 226,
 229, 238
 WCT. *See* Water Cherenkov telescope (WCT)
 Whiffletree, 109, 126
 White pupil, 596–599, 602, 612, 613
 Wilkinson microwave anisotropy probe (WMAP),
 435, 437, 444, 447, 450, 451, 454, 467,
 468, 471–475
 Wind-induced deformation, 305, 306
 Wing buffeting, 6, 25, 35, 39
 WMAP. *See* Wilkinson microwave anisotropy probe
 (WMAP)

Z

Zernike mode, 26
 Zernike polynomials, 26, 27, 107,
 117, 127
 Zernike/Van Cittert Theorem, 245, 246

